IB/05/51057

Europäisches Patentamt | European Patent Office | Office européen des brevets

Bescheinigung | Certificate | Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. | Patent application No. | Demande de brevet n°

04101405.1

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

R C van Dijk

Anmeldung Nr:
Application no.: 04101405.1 ✓
Demande no:

Anmeldetag:
Date of filing: 05.04.04 ✓
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

Koninklijke Philips Electronics N.V.
Groenewoudseweg 1
5621 BA Eindhoven
PAYS-BAS

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se referer à la description.)

Multi-channel parametric audio coding

In Anspruch genommene Priorität(en) / Priority(ies) claimed /Priorité(s) revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/
Classification internationale des brevets:

H04N7/64

Am Anmeldetag benannte Vertragstaaten/Contracting states designated at date of filing/Etats contractants désignées lors du dépôt:

AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IT LU MC NL
PL PT RO SE SI SK TR LI

Multi-channel parametric audio coding

This document gives a technical description of a multi-channel parametric audio coding system as developed by Philips. The goal of this system is to describe an m-channel signal by an n-channel signal, with $n<m$, and parameters describing a spatial image in order to reconstruct the m-channel signal. Although the techniques described in this document could be extended to coding any m to any n channels, the embodiments described in this document is to provide a technical description of coding 5(.1) to 2 or 5(.1) to 1 channel coding. The extension ".1" denotes the presence of an LFE channel. Furthermore, it is assumed that when reproducing the multi-channel signals, a typical loudspeaker setup is used consisting of a Left front (Lf), a Right front (Rf), a Centre (Cf), a Left surround (Ls), a Right surround (Rs) and optionally a low-frequency effects (LFE) speaker.

HIGH LEVEL DESCRIPTION

Figure 1 shows a general block diagram of the multi-channel parametric encoder. The multi-channel input signal consisting of the five channels Lf, Rf, Cf, Ls, Rs and the optional LFE channel are analyzed resulting in a set of parameters describing the spatial image. Depending on the configuration either a mono down-mix channel M is generated or a stereo down mix consisting of the left and right channels Ld and Rd is generated. This mono (M) or stereo (Ld, Rd) signal is then encoded using a conventional mono or stereo audio encoder respectively. The bit stream resulting from this encoding process is merged with a bit stream derived from the coded spatial parameters preferably in a backwards compatible fashion.

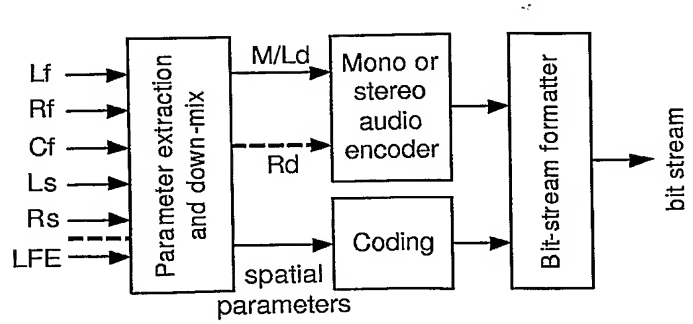

Figure 1: Block diagram of generalized multi-channel parametric encoder.

A block diagram of the corresponding multi-channel parametric decoder is given in Figure 2. First the bit stream is de-multiplexed resulting in a (backwards compatible) bit stream for the mono or stereo audio decoder and a spatial parameter bit-stream. The mono or stereo decoder then reconstructs the coded mono down-mix signal M' or the stereo down-mix signal (Ld',Rd') respectively. Concurrently, the spatial parameters are decoded. Finally the multi-channel signal is reconstructed by imposing the spatial parameters onto the down-mix channel(s).

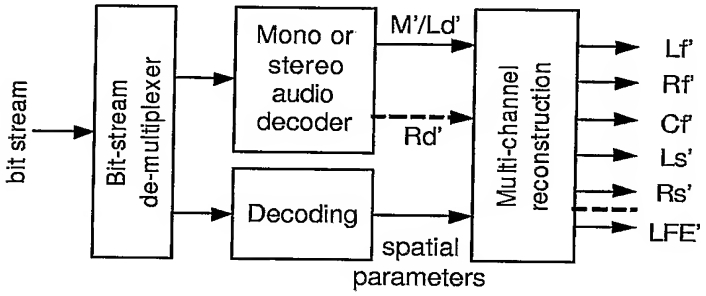

Figure 2: Block diagram of generalized multi-channel parametric decoder.

DETAILED TECHNICAL DESCRIPTION

Time/frequency transform

The multichannel analysis and reconstruction blocks require spatial analysis and synthesis to be performed on individual time/frequency tiles. Therefore, a time/frequency transform is required with the following prerequisites:

The transform is preferably complex, to enable measurement and modification of (relative) phase values between input and output channels;

The transform should be oversampled, to avoid aliasing distortion which would result from time and frequency dependent changes in a critically-sampled system;
The frequency resolution should be non-uniform according to the frequency resolution of the human auditory system;

The time resolution is generally rather low, except in the case of transients.

A generalized block diagram of a spatial encoder is shown in Figure 3. A multi-channel input signal is first transformed to the frequency domain. Subsequently, a downmix and spatial parameters are generated. The downmixed signals are subsequently transformed to the time domain. The decoder basically performs the inverse process.

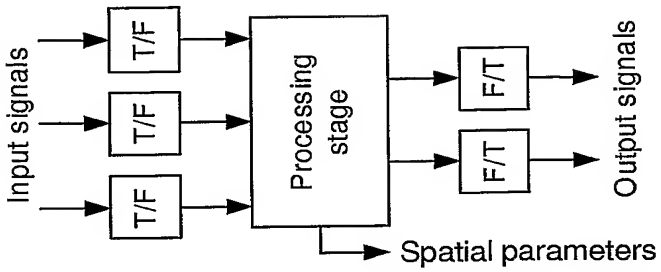

Figure 3: Block diagram of generalized encoder processing stage.

Currently, two time/frequency transforms are used which meet the prerequisites mentioned above. The first transform comprises time-domain segmentation followed by FFTs. All input signals are segmented and transformed by means of the Discrete Fourier Transform (DFT):

$$X_{i,l}[k] = \sum_{n=0}^{N/2} x_i[n + l \cdot h] \cdot h_a[n] \cdot \exp\left(-j\frac{2\pi kn}{N}\right),$$

with $x_i[n]$ the i^{th} time domain input signal, $h_a[n]$ the analysis window, l the frame index, h the frame update in samples, N the DFT length and $X_{i,l}[k]$ the DFT with frequency index k. At the output of the multi-channel encoder/decoder the processed signals $y_i[n]$ are transformed back to the time domain by means of an inverse DFT:

$$y_{i,l}[n] = 2 \cdot h_s[n] \cdot \Re\left\{\sum_{n=0}^{N-1} Y_{i,l}[k] \cdot \exp\left(j\frac{2\pi kn}{N}\right)\right\},$$

with $Y_{i,l}[k]$ the (zero-padded) DFT representation, $y_{i,l}[n]$ the time-domain segment corresponding to frame l and $h_s[n]$ the synthesis window. The resulting output signal $y_i[n]$ is obtained by overlap-add of the segments $y_{i,l}[n]$.

The second transform which is of particular interest for memory and computational complexity reasons is a complex-exponential modulated filter bank.

In the following sections, it is assumed that individual time/frequency tiles of all input channels are available for processing at both the encoder and decoder side.

System based on Stereo downmix

Encoder

The aim of this encoder is to represent a 5.1-channel input signal as a backwards-compatible stereo signal (i.e., with a spatial representation which resembles the

5.1 channel reconstruction as good as possible), combined with spatial parameters that enable reconstruction of a 5.1-channel output that resembles the original 5.1 input signals from a perceptual point of view. The structure of this system is depicted in Fig. 4.

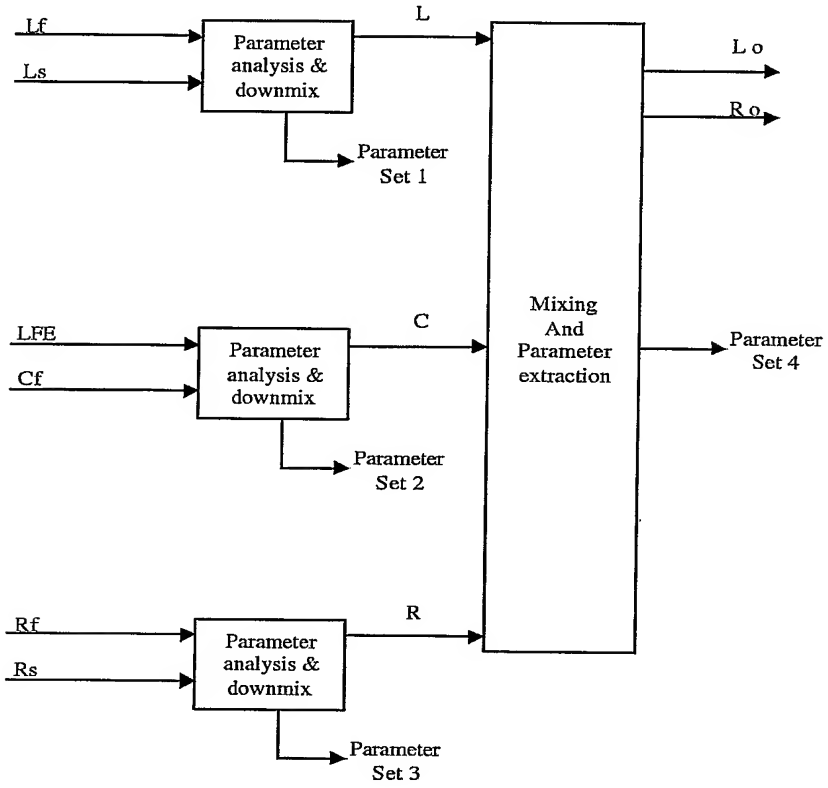

Figure 4: Block diagram of an encoder providing a stereo downmix.

The encoder consists of three parallel stages which all convert a stereo input signal to a mono signal, combined with parameter extraction which represents the spatial cues between the respective input signals. Each of these three parallel blocks computes (assuming input signals X_1 and X_2, and output signal X_m):

The ratio of the powers of corresponding time/frequency tiles of the input signals (which will be denoted 'Interchannel Level Difference', or ILD), given by:

$$ILD = 10\log 10\left(\frac{\sum_k X_1[k]X_1^*[k]}{\sum_k X_2[k]X_2^*[k]}\right).$$

The average phase difference (or the phase difference which maximizes the correlation between the input signals), which is referred to as 'Interchannel Phase

Difference', or IPD, given by:

$$IPD = \angle\left(\sum_k X_1[k]X_2^*[k]\right).$$

The coherence (ICC), which is the normalized cross-correlation between the input signals given the IPD value as in (2):

$$ICC = \left|\left(\frac{\sum_k X_1[k]X_2^*[k]}{\sqrt{\sum_k X_1[k]X_1^*[k]\sum_k X_2[k]X_2^*[k]}}\right)\right|.$$

Each parallel processing block generates a parameter set which comprises all or a selected number of these parameters for each time/frequency tile (depending on the desired parameter bitrate, some parameters are not transmitted). Besides parameters, each parallel stage also computes a single output signal, which is a linear (complex) combination of the two input signals:

$X_m[k] = w_1 X_1[k] + w_2 X_2[k]$,

with w_1 and w_2 complex weights which depend on the extracted parameters (w_i = f(IID, IPD, ICC)). Preferably each time/frequency tile of X_m has a power that is equal to the sum of the powers of the input signals X_1 and X_2.

A fourth parameter which is required for reconstruction of phase differences in the encoder is the Overall Phase Difference (or OPD), which is defined as the average phase difference between the first input signal and the mono output signal.

The next stage of the encoder performs a down-mix from three (L, C, R) to two down-mix channels (Lo, Ro), combined with corresponding spatial parameters. Each down-mix channel is a linear combination of the input signals L, R and C:

$L_0[k] = \alpha_1 L[k] + \alpha_2 R[k] + \alpha_3 C[k]$,
$R_0[k] = \beta_1 L[k] + \beta_2 R[k] + \beta_3 C[k]$.

The parameters α_i and β_i are chosen such that a good stereo image of the stereo signal consisting of $L_0[k]$ and $R_0[k]$ is obtained. One of the prerequisites for a good stereo image is that a_3 equals $ß_3$.

At the decoder, channels L, R and C are predicted using the two down-mix channels Lo and Ro as follows:

$$\hat{L}[k] = C_{1,L} L_0[k] + C_{2,L} R_0[k],$$
$$\hat{R}[k] = C_{1,R} L_0[k] + C_{2,R} R_0[k],$$
$$\hat{C}[k] = C_{1,C} L_0[k] + C_{2,C} R_0[k].$$

To this end, parameters $C_{1,Z}$ and $C_{2,Z}$ (for Z = L, R or C) are computed at the encoder and sent to the decoder.

A minimal Euclidian norm of the difference of signal $Z[k]$ and its estimation $\hat{Z}[k]$ is used as optimization criterion to find the parameters $C_{1,Z}$ and $C_{2,Z}$. The square of the Euclidian norm of the difference of the original input channel $Z[k]$ and its estimation at the decoder $\hat{Z}[k]$ can be written as:

$$\sum_k \left| Z[k] - \hat{Z}[k] \right|^2,$$

with

$$\hat{Z}[k] = C_{1,Z} L_0[k] + C_{2,Z} R_0[k].$$

Minimization of $\sum_k \left| Z[k] - \hat{Z}[k] \right|^2$ leads to the following expressions:

$$C_{1,Z} = \frac{< L_0[k], Z[k] >^* \left\| R_0[k] \right\|^2 - < R_0[k], Z[k] >^* < L_0[k], R_0[k] >^*}{\left\| L_0[k] \right\|^2 \left\| R_0[k] \right\|^2 - \left| < L_0[k], R_0[k] > \right|^2},$$
$$C_{2,Z} = \frac{< R_0[k], Z[k] >^* \left\| L_0[k] \right\| - < L_0[k], Z[k] >^* < L_0[k], R_0[k] >}{\left\| L_0[k] \right\|^2 \left\| R_0[k] \right\|^2 - \left| < L_0[k], R_0[k] > \right|^2},$$

with

$$\left\| A[k] \right\|^2 = \sum_k \left| A[k] \right|^2,$$
$$< A[k], B[k] >= \sum_k A[k] B^*[k].$$

For the coefficients $C_{1,Z}$ and $C_{2,Z}$ the following relations can be derived:

$$\alpha_1 C_{1,L} + \alpha_2 C_{1,R} + \alpha_3 C_{1,C} = 1,$$
$$\beta_1 C_{2,L} + \beta_2 C_{2,R} + \beta_3 C_{2,C} = 1,$$
$$\alpha_1 C_{2,L} + \alpha_2 C_{2,R} + \alpha_3 C_{2,C} = 0,$$
$$\beta_1 C_{1,L} + \beta_2 C_{1,R} + \beta_3 C_{1,C} = 0.$$

Having 6 variables ($C_{1,L}$, $C_{2,L}$, $C_{1,R}$, $C_{2,R}$, $C_{1,C}$ and $C_{2,C}$) and at the same time 4 relations between these parameters, only 2 parameters need to be sent to the decoder. $C_{1,L}$ (henceforth notated by β) and $C_{2,R}$ (henceforth notated by γ) are transmitted to the decoder because of their similar statistical distributions.

Before the actual downmix is applied (i.e., L, C and R are combined to Lo and Ro), the signals L, C and R are preferably pre-conditioned by first applying a phase shift to L, R, and/or C to assure that the signal pair L and C on the one hand and R and C on the other hand have a non-negative correlation to minimize energy loss by the downmix process. The applied phase shifts can be transmitted without any bit-rate costs by superimposing these phase shifts on the OPD values resulting from the individual two-to-one stages.

If no LFE channel is present at the encoder input, the LFE channel can be considered as containing zeros only and all related processing steps can be discarded. In that case, parameter set 2 as shown in Figure 4 is irrelevant and does not have to be transmitted.

Decoder

The decoder basically performs the reverse process as depicted in Figure 4. In a first stage, the stereo input signal (Lo, Ro) is converted to a three-channel signal (L, C, R) based on parameter set 4. The upmix, based on the transmitted parameters ß and γ, is performed as follows:

$$L[k] = \beta L_0[k] + (\gamma - 1)R_0[k]$$
$$R[k] = (\beta - 1)L_0[k] + \gamma R_0[k]$$
$$C[k] = (1-\beta)L_0[k] + (1-\gamma)R_0[k]$$

If a 3-channel reconstruction is desired, the decoding process is finished.

For e.g. a 3.1, 5 or 5.1-channel reconstruction, the mono signals L, C, and R are subdivided into two-channel signals, based on the spatial parameters corresponding to each signal. The general structure of the mono-to-stereo upmix block is given in Figure 5.

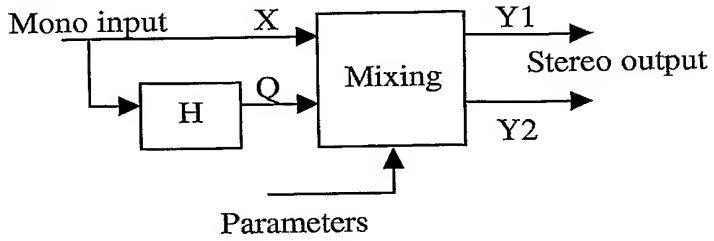

Figure 5: Generalized structure of the mono-to-stereo upmix stage.

A local copy of the mono input signal X is fed through a decorrelation filter H. This filter can be implemented as a (frequency-dependent) delay or reverberation module. The mono input signal X and its decorrelated copy Q are subsequently combined in a mixing stage to form the stereo output signal (Y_1, Y_2):

$$\begin{bmatrix} Y_1[k] \\ Y_2[k] \end{bmatrix} = \begin{bmatrix} \exp(jOPD_L) & 0 \\ 0 & \exp(jOPD_L - IPD_L) \end{bmatrix} \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \cos(\tau) & 0 \\ 0 & \sin(\tau) \end{bmatrix} \begin{bmatrix} X[k] \\ Q[k] \end{bmatrix}$$

with

.

$$\tau = \arctan\left(\frac{1-\sqrt{\mu}}{1+\sqrt{\mu}}\right),$$

$$\alpha = \frac{1}{2}\arctan\left(\frac{2gICC}{g^2-1}\right),$$

$$\mu = 1+\frac{4ICC^2-4}{\left(g+1/g\right)^2},$$

$$g = 10^{IID/20}.$$

If desired, this process is repeated for all three signals L, C, and R, resulting in a 5.1 (Lf, Rf, Ls, Rs, Cf, LFE) signal. It should be noted that the decorrelation filter H can be different for different mono-to-stereo processing blocks.

Enhancement methods

The backwards-compatible stereo downmix will usually result in a significantly-reduced spatial image compared to the original 5-channel reconstruction. It is possible to enhance the stereo downmix in such a way that the perceived spatial image of the downmix resembles the 5-channel reconstruction more closely by introducing virtual surround loudspeakers (or sometimes denoted as 'stereo widening' algorithm). The generation of virtual surround loudspeakers is based on cross-talk cancellation principles. Various methods for stereo-widening are currently available. However, these systems have an important drawback: if they are applied on a stereo downmix they result in a widening of the complete sound stage instead of a widening of the surround signals only. To overcome this drawback, we propose a cross-talk cancellation algorithm that is applied on the stereo downmix, in which the amount of cross-talk cancellation depends on the spatial encoding parameters. Using this method, (1) only signal parts that would have been reproduced by surround loudspeakers in a 5-channel setup are processed, and (2) the cross-talk cancellation algorithm can be inverted, which is a requirement for high-quality 5-channel reconstruction.

System based on mono-downmix

Approach 1

Encoder

A straightforward approach for a 5.1-to-1 encoder is depicted in Figure 6. The encoder consists of subsequent stereo-to-mono analysis blocks, which are identical to those used in Figure 4.

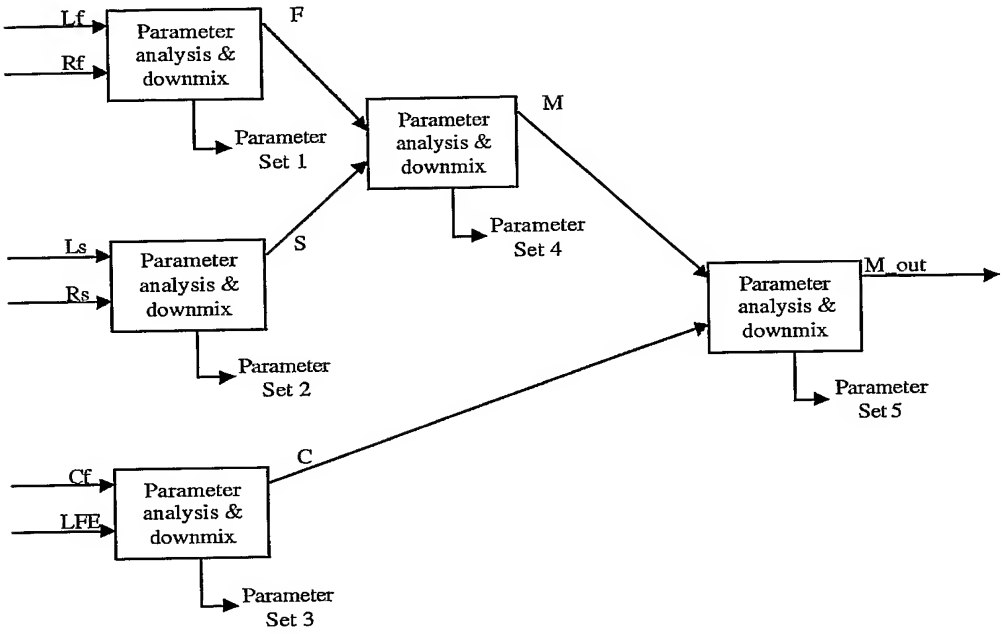

Figure 6: Structure of 5.1-to-one encoder.

This approach reduces the number of input channels pair-wise using stereo-to-mono blocks. Each block generates a mono signal and spatial parameters, of which a selection is transmitted. A recommended parameter selection includes:

Parameter set 1:

IID, ICC for each time/frequency tile;

IPD, OPD for time/frequency tiles up to about 2 kHz.

Parameter set 2:

IID, ICC for each time/frequency tile;

IPD, OPD for time/frequency tiles up to about 2 kHz.

Parameter set 3:

IID only for time/frequency tiles up to about 150 Hz.

Parameter set 4:

IID, ICC for each time/frequency tile;

Optionally: IPD, OPD for time/frequency tiles up to about 2 kHz.

Parameter set 5:

IID for each time/frequency tile.

Decoder

A corresponding decoder consists of the reverse process as depicted in Figure 6. The individual building blocks are equal to those described in detail in Section 0.

Approach 2

Instead of performing a spatial analysis and downmix on a pair-wise basis, this approach performs the downmix in a single stage based on multidimensional Principal Component Analysis (PCA). Assuming 5 complex-valued input signals X_i, the complex-valued covariance matrix S_{ij} of the incoming signals is given by:

$$S_{ij} = \sum_k X_i X_j^c .$$

Note that $S_{ij} = S_{ji}^c$. In principle the non-diagonal elements of the covariance matrix (i.e., i not equal to j) are complex. However, it is possible to extract and transmit the complex angles of S_{ij} separately, resulting in a covariance matrix containing elements equal or larger than zero only. We will describe a real-valued covariance matrix from this point on.

It is assumed that the input signals X_i consist of a rotation (R) of a set of orthogonal signals Y_i (i.e., $R^{-1} = R^T$)

$$X = RY .$$

Given the fact that the individual signals of Y_i are orthogonal, the covariance matrix of Y, S_y, is diagonal. Consequently, the covariance matrix of X, S_x, can be written as:

$$S_x = RS_y R^T .$$

Given the diagonal matrix S_y, the above expression equals the eigenvalue/eigenvector decomposition of S_x. This means that the rotation matrix R and the

eigenvalues (which are equal to the energies of the orthogonal signals of Y) can be obtained by an eigenvalue/eigenvector decomposition of the covariance matrix S_x.

The general encoder approach then consists of:

computing and applying (relative) phase parameters of the input signals in order to result in a covariance matrix which contains smaller imaginary parts; a possible technique is applying complex principal component analysis;

computing the eigenvalue/eigenvector decomposition of the complex covariance matrix, resulting from the original or modified input signals;

(inverse) rotation of the input signals to obtain the principal components;

transmission of the first principal component only (i.e., the component corresponding to the largest eigenvalue) as mono output signal

transmission of:

either the rotation matrix R (e.g. in terms of its angular components) and the (relative) powers of each principal component, or

the (real- or complex valued) covariance matrix itself in terms of IPDs, correlations and power ratios;

the relative complex phase angles;

preferably an overall phase shift of the principal component signal.

The decoder consists of the following steps:

computing the rotation matrix R (either from the direct transmission or an eigenvalue decomposition of the transmitted covariance matrix;

computing of a set of orthogonal signals by means of decorrelating the mono input signal;

scaling the orthogonal signals by means of the transmitted (relative) signal powers;

generation of a five-channel output by rotating the orthogonal signals;

re-instating the complex phase relationships based on transmitted (relative, IPD and overall, OPD) phase information.

Approach 3

The five-dimensional PCA as described as Approach 2 is relatively complex compared to the cascaded approach described in Approach 1 especially since it is not possible to derive the eigenvectors analytically. As a compromise between both approaches the cascaded approach could be combined with three-dimensional PCA. This is illustrated in Figure 7.

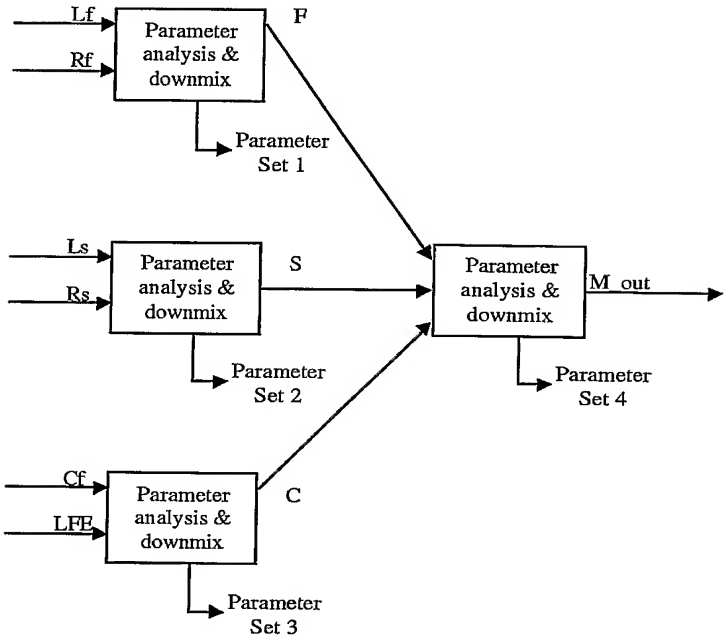

Figure 7: Hybrid encoder approach.

The front (F), surround (S) and centre (C) channels are derived similar to Approach 1. Consequently these signals are used as input to a three-dimensional PCA procedure resulting in a 3x3 rotation matrix R. The same procedure for encoding and decoding as described in Approach 2 can be used.

Alternatively, the three channels input to the three-dimensional PCA procedure could consist of

Downmix of Lf and Ls

Downmix of Rf and Rs

Downmix of Cf and LFE

Residual-coding extensions

Most encoder processing blocks reduce the number of input channels, combined with a parameterization of the relations between the original input channels. In principle, each reduction in number of channels results in an output signal, but also in a complementary residual signal (or more residual signals), which would in principle allow perfect reconstruction of the original input signals. In the decoder, these residual signals are artificially regenerated by decorrelation filters. It should be noted, however, that it can occur for certain time/frequency tiles that this synthetic residual signal is inappropriate for a perceptually high-quality decoder output signal. For these time/frequency tiles, it is possible

to encode and transmit the actual residual signal from the encoder, while for remaining time/frequency tiles, the decoder can use the synthetic residual signal. The scalability options of this approach are of interest; the encoded residual signals can simply be removed from the bitstream. In that case, the decoder will default to its synthetic residual signal. This approach can be used for three-to-one, three-to-two, as well as two-to-one processing blocks.

Bit-stream syntax

As already mentioned on page 1, the basic techniques used for multi-channel parametric coding can be applied to code different multi-channel signal configurations. This is reflected in the proposed bit-stream syntax. Some example possibilities:

encode an Lf, Rf, Ls, Rs (4 channels) multi-channel signal into a mono or stereo-compatible signal,

encode a L(f),C,R(f),LFE (3.1 channels) multi-channel signal into a mono or stereo-compatible signal or

encode a Lf,C,Rf,Ls,Rs,LFE (5.1 channels) multi-channel signal into a L(f), C, R(f) multi-channel compatible signal.

The bit-stream should be very flexible for decoding subsets of the multi-channel signal. Consider the situation where the original signal consisted of a 5.1 multi-channel signal encoded into a stereo signal using a similar structure as depicted in Figure 4. Furthermore, assume that reconstruction is required for a 3.1 setup, consisting of a Left (front), Right (front), Centre (front) and an LFE speaker. By only decoding using parameter sets 2 and 4 the 3.1 channel signal is already obtained. Hence, it is not necessary to further decode the surround channels.

In the bit-stream syntax the flexibility described above is obtained by explicitly defining the channel configuration of each encoder/decoder step in the mc_channel_config element. In this element it is described which parameters belong to which (intermediate) input and (intermediate) output channels (see Table 16).

Consider the very simple case where the input signal consists of the mono signal M and the output signal consists of the mono signal M' and the signal LFE. If the frequency range of the LFE signal is only limited with respect to the bandwidth of the mono signal M, which is typically the case, only for the lower part of the frequency range the signals M' and LFE are obtained. For the higher part of the frequency range the signal M' is simply obtained as M'=M. A similar situation might occur in a 1 to 3 decoding step. Consider the case where the input signal consists of mono signal M and the output signal consists of

the left signal L(f), the right signal R(f) and the LFE signal. For the frequency range for which the LFE has been included, parameters for all three signals are included, which may e.g. consist of IID values between L and R and between L and LFE and coherence values between L and R, L and LFE *and* R and LFE (3 sets!). For the higher frequency range, the parameters will then only consist of IID values between L and R and coherence values between L and R. For this frequency range the decoder operates in a 1 channel to 2 channels decoding mode.

For the elements data1to3() two alternatives are given, the first one uses a representation containing the parameters of the covariance matrix, split into power ratios and coherence values. The second alternative uses an angular representation (e.g. as defined by Euler) of the rotation matrix R.

Table 1 – Syntax of mc_data()

Syntax	Num. bits	Mnemonic
mc_data(){		
if (**mc_channel_config**==1) {	**1**	**uimsbf**
mc_channel_config()		
}		
for (s=0; s<nr_steps; s++) {		Note 1
switch(method[s]) {		
case 0:		
data1to2()		
break;		
case 1:		
data2to3()		
break;		
case 2:		
data1to3()		
break;		
case 3:		
data1to5()		
break;		
case 4:		
data1to51()		
break;		
}		
}		
}		
Note 1: nr_steps is defined in the mc_channel_config() element. Hence, for the first instance of mc_data() mc_channel_config should be set to %1.		

Table 2 – Syntax of mc_channel_config()

Syntax	Num. bits	Mnemonic
mc_channel_config() {		
nr_downmix_ch_coded	**1**	Note 1
nr_output_ch_coded	**3**	Note 2
nr_steps = 0;		
nr_dec_ch = nr_downmix_ch;		
while (nr_dec_ch < nr_output_ch) {		
for (ch=0; ch<method_in_ch[nr_steps]; ch++) {		
input_ch[nr_steps,ch]	**3**	Note 3
}		
for (ch=0; ch<method_out_ch[nr_steps]; ch++) {		
output_ch[nr_steps,ch]	**3**	Note 3
}		
method[nr_steps]	**3**	
nr_dec_ch += nr_increase_ch[method[nr_steps]];		
nr_steps += 1		
}		
num_env_default	**2**	Table 15
}		

Note 1: nr_downmix_ch = nr_downmix_ch_coded + 1
Note 2: nr_output_ch = nr_downmix_ch + nr_output_ch_coded;
Note 3: See Table 16. LFE can never be input channel!

Table 3 – Syntax of data1to2

Syntax	Num. bits	Mnemonic
data1to2{		
if (**data1to2_header**) {	**1**	
if (**enable_iid**) {	**1**	**uimsbf**
iid_mode	**??**	
}		
if (**enable_icc**) {	**1**	**uimsbf**
icc_mode	**??**	
}		
if (**enable_ipdopd**) {	**1**	**uimsbf**
ipdopd_mode	**??**	
}		
freq_res	**2**	Table 17
if (**var_framing**) {	**1**	**uimsbf**
for (e=0; e< **num_env**; e++) {	**2**	**uimsbf**
env_pos[e]	**5**	**uimsbf**
}		
}		
for (ch=0; ch<method_out_ch[nr_steps]; ch++) {		
if (output_ch[nr_steps,ch]==6) {		
nr_bands_coded	**??**	Note 1
}		
}		
}		
if (enable_iid) {		
for (e=0; e<num_env; e++) {		
iid_dt[e]	**1**	**uimsbf**
iid_data(e,nr_bands_coded)		
}		
}		
if (enable_icc) {		
for (e=0; e<num_env; e++) {		
icc_dt[e]	**1**	**uimsbf**
icc_data(e,nr_bands_coded)		
}		
}		

if (enable_ipdopd) {		
for (e=0; e<num_env; e++) {		
ipd_dt[e]	**1**	**uimsbf**
ipd_data(e,nr_ipdopd_coded)		
opd_dt[e]	**1**	**uimsbf**
opd_data(e,nr_ipdopd_coded)		
}		
}		
}		
Note 1: In case one of the output channels is the LFE channel only a part of the signal is coded, denoted with nr_bands_coded.		

Table 4 – Syntax of dta2to3

Syntax	Num. bits	Mnemonic
data2to3{		
if (**data2to3_header**) {	**1**	
if (**enable_beta**) {	**1**	**uimsbf**
beta_mode	**??**	
}		
if (**enable_gamma**) {	**1**	**uimsbf**
gamma_mode	**??**	
}		
if (**enable_opd**) {	**1**	**uimsbf**
opd_mode	**??**	
}		
freq_res	**2**	
if (**var_framing**) {	**1**	**uimsbf**
for (e=0; e<**num_env_lcr**; e++) {	**2**	**uimsbf**
env_pos_lcr[e]	**5**	**uimsbf**
}		
} else {		
num_env_lcr = num_env_frame		
}		
for (ch=0; ch<method_out_ch[nr_steps]; ch++) {		
if (output_ch[nr_steps,ch]==6) {		
nr_bands_coded	**??**	Note 1
}		
}		
}		
if (enable_beta) {		
for (e=0; e<num_env; e++) {		
beta_dt[e]	**1**	**uimsbf**
beta_data(e,nr_bands_coded)		
}		
}		
if (enable_gamma) {		
for (e=0; e<num_env; e++) {		
gamma_dt[e]	**1**	**uimsbf**
gamma_data(e,nr_bands_coded)		
}		
}		

if (enable_opd) {
for (e=0; e<num_env; e++) {
opd_dt[e]
opd_data(e,nr_opd_coded)
}
}
	1	**uimsbf**

Note 1: In case one of the output channels is the LFE channel only a part of the signal is coded, denoted with nr_bands_coded.		

Table 5 - Syntax of data1to3 (Alternative A)

Syntax	Num. bits	Mnemonic
data1to3{		
if (**data1to3_header**) {	**1**	
if (**enable_iid**) {	**1**	**uimsbf**
iid_mode	**??**	
}		
if (**enable_icc**) {	**1**	**uimsbf**
icc_mode	**??**	
}		
if (**enable_ipdopd**) {	**1**	**uimsbf**
ipdopd_mode	**??**	
}		
freq_res	**2**	Table 17
if (**var_framing**) {	**1**	**uimsbf**
for (e=0; e< **num_env**; e++) {	**2**	**uimsbf**
env_pos[e]	**5**	**uimsbf**
}		
}		
for (ch=0; ch<method_out_ch[nr_steps]; ch++) {		
if (output_ch[nr_steps,ch]==6) {		
nr_bands_coded_set2	**??**	Note 1
}		
}		
}		
if (enable_iid) {		
for (e=0; e<num_env; e++) {		
iid_dt[e]	**1**	**uimsbf**
iid_data(e,nr_bands_coded)		
iid_dt[e]	**1**	**uimsbf**
iid_data(e,nr_bands_coded_set2)		
}		
}		
if (enable_icc) {		
for (e=0; e<num_env; e++) {		
icc_dt[e]	**1**	**uimsbf**
icc_data(e,nr_bands_coded)		
icc_dt[e]	**1**	**uimsbf**

icc_data(e,nr_bands_coded_set2) **icc_dt**[e]	**1**	**uimsbf**

icc_data(e,nr_bands_coded_set2)
}
}
if (enable_ipdopd) {
nr =
min(nr_bands_coded_set2,nr_ipdopd_coded);
for (e=0; e<num_env; e++) {		

ipd_dt[e] ipd_data(e,nr_ipdopd_coded)	**1**	**uimsbf**
ipd_dt[e] ipd_data(e,nr)	**1**	**uimsbf**
opd_dt[e] opd_data(e,nr)	**1**	**uimsbf**
} } }		

Note 1: In case one of the output channels is the LFE channel only a part of the signal is decoded to three channels, denoted with nr_bands_coded_set2. The rest of the signal is decoded to two channels, i.e., nr_bands_coded_set2 is set to nr_bands_coded. Note that for the bands decoded to three channels there exist two sets of IIDs, three sets of ICCs, two sets of IPDs and one set of OPDs.

Table 6 - Syntax of data1to3 (Alternative B)

Syntax	Num. bits	Mnemonic
data1to3{		
if (**data1to3_header**) {	**1**	
if (**enable_angles**) {	**1**	**uimsbf**
angle_mode	**??**	
}		
if (**enable_ipdopd**) {	**1**	**uimsbf**
ipdopd_mode	**??**	
}		
freq_res	**2**	Table 17
if (**var_framing**) {	**1**	**uimsbf**
for (e=0; e< **num_env**; e++) {	**2**	**uimsbf**
env_pos[e]	**5**	**uimsbf**
}		
}		
for (ch=0; ch<method_out_ch[nr_steps]; ch++) {		
if (output_ch[nr_steps,ch]==6) {		
nr_bands_coded_set2	**??**	Note 2
}		
}		
}		
if (enable_angles) {		
for (e=0; e<num_env; e++) {		
angle_dt[e]	**1**	**uimsbf**
angle_data(e,nr_bands_coded)		
angle_dt[e]	**1**	**uimsbf**
angle_data(e,nr_bands_coded_set2)		
}		
}		
if (enable_ipdopd) {		

nr = min(nr_bands_coded_set2,nr_ipdopd_coded);
for (e=0; e<num_env; e++) {
ipd_dt[e]
	1	**uimsbf**

ipd_data(e,nr_ipdopd_coded) **ipd_dt**[e]	**1**	**uimsbf**
ipd_data(e,nr) **opd_dt**[e]	**1**	**uimsbf**
opd_data(e,nr) } } }		
Note 1: In case one of the output channels is the LFE channel only a part of the signal is decoded to three channels, denoted with nr_bands_coded_set2. The rest of the signal is decoded to two channels, i.e., nr_bands_coded_set2 is set to nr_bands_coded. Note that for the bands decoded to three channels there exist two sets of IIDs, three sets of ICCs, two sets of IPDs and one set of OPDs.		

The data1to5() and data1to51() elements are straightforward extensions of the data1to3() element with an increased amount of parameter sets.

Table 7 - Syntax of iid_data()

Syntax	Num. bits	Mnemonic

iid_data(e, nr_iid_par) {
if (iid_dt[e]) {
for (b=0 ; b<nr_iid_par; b++) {
iid_par_dt[e][b] = sa_huff_dec(huff_iid_dt[iid_quant],**bs_codeword**);
}
}
else {
for (b=0 ; b<nr_iid_par; b++) {
iid_par_df[e][b] = sa_huff_dec(huff_iid_df[iid_quant],**bs_codeword**);
}
}
}
	??...??
	
	??...??
		bslbf
		
		bslbf

Table 8 - Syntax of icc_data()

Syntax	Num. bits	Mnemonic

icc_data(e, nr_icc_par) {
if (icc_dt[e]) {
for (b=0 ; b<nr_icc_par; b++) {
icc_par_dt[e][b] = sa_huff_dec(huff_icc_dt,**bs_codeword**);
}
}
else {
for (b=0 ; b<nr_icc_par; b++) {
icc_par_df[e][b] = sa_huff_dec(huff_icc_df,**bs_codeword**);
}
}
}
	??...??
	
	??...??
		bslbf
		
		bslbf

Table 9 - Syntax of ipd_data()

Syntax	Num. bits	Mnemonic

ipd_data(e, nr_ipd_par) {
if (ipd_dt[e]) {
for (b=0 ; b<nr_ipdopd_par; b++)
{
ipd_par_dt[e][b] = sa_huff_dec(huff_ipd_dt,**bs_codeword**);
}
}
else {
for (b=0 ; b<nr_ipdopd_par; b++)
{
ipd_par_df[e][b] = sa_huff_dec(huff_ipd_df,**bs_codeword**);
}
}
}
	??...??
	
	??...??
		bslbf
		
		bslbf

5

Table 10 - Syntax of opd_data()

Syntax	Num. bits	Mnemonic

opd_data(e, nr_opd_par) {
if (opd_dt[e]) {
for (b=0 ; b<nr_ipdopd_par;
b++) {
opd_par_dt[e][b] =
sa_huff_dec(huff_opd_dt,**bs_codeword**);
}
}
else {
for (b=0 ; b<nr_ipdopd_par;
b++) {
opd_par_df[e][b] =
sa_huff_dec(huff_opd_df,**bs_codeword**);
}
}
}
	??...??
	
	??...??
		bslbf
		
		bslbf

Table 11 - Syntax of beta_data()

Syntax	Num. bits	Mnemonic

beta_data(e, nr_beta_par) {
if (beta_dt[e]) {
for (b=0 ; b<nr_beta_par; b++) {
beta_par_dt[e][b] =
sa_huff_dec(huff_beta_dt,**bs_codeword**);
}
}
else {
for (b=0 ; b<nr_beta_par; b++) {
beta_par_df[e][b] =
sa_huff_dec(huff_beta_df,**bs_codeword**);
}
}
}
	??...??
	
	??...??
		bslbf
		
		bslbf

Table 12 - Syntax of gamma_data()

Syntax	Num. bits	Mnemonic

gamma_data(e, nr_gamma_par) {
if (gamma_dt[e]) {
for (b=0 ; b<nr_gamma_par;
b++) {
gamma_par_dt[e][b] =
sa_huff_dec(huff_gamma_dt,**bs_codeword**);
}
}
else {
for (b=0 ; b<nr_gamma_par;
b++) {
gamma_par_df[e][b] =
sa_huff_dec(huff_gamma_df,**bs_codeword**);
}
}
}
	??...??
	
	??...??
		bslbf
		
		bslbf

Table 13 - Syntax of angle_data()

Syntax	Num. bits	Mnemonic

angle_data(e, nr_angle_par) {
for (a=0 ; a<nr_angles ; a++)
if (angle_dt[a,e]) {
for (b=0 ;
b<nr_angle_par; b++) {
angle_par_dt[a][e][b] =
sa_huff_dec(huff_gamma_dt,**bs_codeword**);
}
}
else {
for (b=0 ;
b<nr_angle_par; b++) {
angle_par_df[a][e][b] =
sa_huff_dec(huff_gamma_df,**bs_codeword**);
}
}
}
}
	??...??
	
	??...??
		Note 1
		
		bslbf
		
		bslbf

Note 1: nr_angles follows from used method!		

5

Table 14: Dependencies of variable method

method	description	method_in_ch	method_out_ch	nr_increase_ch
0	1 to 2	1	2	1
1	2 to 3	2	3	1
2	1 to 3	1	3	2
3	1 to 5	0 (fixed)	0 (fixed)	4
4	1 to 5.1	0 (fixed)	0 (fixed)	5
5	reserved	-	-	-
6	reserved	-	-	-
7	reserved	-	-	-

Table 15: num_env as a function of num_env_default

num_env_default	num_env
0	0
1	1
2	2
3	4

Table 16: Channel description

input_ch / output_ch	description	abbreviation
0	mono	M
1	left (front)	L(f)
2	right (front)	R(f)
3	left surround	Ls
4	right surround	Rs
5	centre (or front)	C (F)
6	low frequency effects	LFE
7	surround	S

Table 17: nr_bands and nr_bands_coded as a function of freq_res

freq_res	nr_bands	nr_bands_coded (per default)
0	10	10
1	20	20
2	34	34
3	reserved	-

Annex A Parametric multichannel audio coder with 2, 3, 4 and 5 channel playback compatibility

This invention is a multi-channel extension of the basic principle described in WO03/090208-A1. The 5-channel content is downmixed into 2 channels combined with a small amount of parametric overhead which enables 5-channel reconstruction at the decoder side. Moreover, 2, 3, and 4-channel reproduction are also supported.

The important features of the proposed coder are:

transmission of two audio channels (which can be encoded using an arbitrary stereo audio codec) and are preferably obtained using principal component analysis on the left-front and left-rear pair on the one hand, and using a separate principal component analysis on the right-front and right-rear signal pair;

- transmission of parametric overhead, which includes:
- inter-channel level differences between left-front and left-rear channels;
- inter-channel level differences between right-front and right-rear channels;
- inter-channel coherence values between left-front and left-rear channels;
- inter-channel coherence values between right-front and right-rear channels;
- the power ratio between the centre channel and the sum of the powers of left-front, left-rear, right-front and right-rear channels

Additionally, the inter-channel phase differences and overall phase differences between left-front and left-rear on the one hand, and right-front and right-rear on the other hand, may also be included in the parametric bit stream.

The parameters described above are typically analyzed as a function of time and frequency (i.e., for a set of time/frequency tiles).

Encoder

Assume a five-channel audio signal $l_f[n]$, $l_r[n]$, $r_f[n]$, $r_r[n]$, $c[n]$, which describe the discrete time-domain waveforms for the left-front, left-rear, right-front, right-rear and centre signals, respectively. These five signals are segmented using a common segmentation, preferably using overlapping analysis windows. Subsequently, each segment is converted to the frequency domain using a complex transform (e.g., FFT). However, complex filter-bank structures may also be appropriate to obtain time/frequency tiles. This process results in segmented, sub-band representations of the input signals (which will be denoted by $L_f[k]$, $L_r[k]$, $R_f[k]$, $R_r[k]$, and $C[k]$ with k denoting frequency index).

As a first step, the relevant parameters between left-front and left-rear are estimated. These parameters include the level difference (IID_L), the (average) phase difference (IPD_L) and the coherence (ICC_L):

$$IID_L = 10\log 10\left(\frac{\sum_k L_f[k]L_f^*[k]}{\sum_k L_r[k]L_r^*[k]}\right)$$

$$IPD_L = \angle\left(\frac{\sum_k L_f[k]L_r^*[k]}{\sqrt{\sum_k L_f[k]L_f^*[k]\sum_k L_r[k]L_r^*[k]}}\right)$$

$$ICC_L = \left|\left(\frac{\sum_k L_f[k]L_r^*[k]}{\sqrt{\sum_k L_f[k]L_f^*[k]\sum_k L_r[k]L_r^*[k]}}\right)\right|$$

This process is repeated for the right-front and right-rear channels, resulting in IID_R, IPD_R, and ICC_R.

The second step consists of a principal component analysis (PCA) of the two signals left-front (L_f) and left-rear (L_r). To be more specific, these two input signals are rotated in order to obtain a dominant ($Y[k]$) and a residual signal ($Q[k]$), using a rotation angle a which maximizes the energy of the dominant signal:

$$\begin{bmatrix} Y[k] \\ Q[k] \end{bmatrix} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} L_f[k].\exp(j(-OPD_L)) \\ L_r[k].\exp(j(-OPD_L + IPD_L)) \end{bmatrix}$$

Here, the angle OPD_L denotes an overall phase rotation angle, while IPD_L ensures maximum phase-alignment of the two signals L_f and L_r. The rotation angle a can be derived from the IID_L and ICC_L parameters following

$$\alpha = \frac{1}{2}\arctan\left(\frac{2gICC_L}{g^2 - 1}\right)$$

with

$$g = 10^{IID_L/20}$$

The signal $Q[k]$ is subsequently discarded, and signal $Y[k]$ is scaled by a scalar β to obtain $L[k]$ in such a way that $L[k]$ has the same power as the power of $Q[k]$ plus the power of $Y[k]$ (i.e., the signal $Q[k]$ is discarded while the loss in signal power is compensated for by scaling of $Y[k]$). It can be shown that the required scale factor β is equal to:

$$\beta = \sqrt{1+\frac{1-\sqrt{\mu}}{1+\sqrt{\mu}}}$$

with

$$\mu = 1+\frac{4ICC_L{}^2-4}{\left(g+1/g\right)^2}$$

This process is also repeated for the right-front and right-rear signal pair, resulting in signal $R[k]$.

The last step entails mixing the centre signal $C[k]$ in both $L[k]$ and $R[k]$, resulting in the stereo output signal pair $L_{OUT}[k]$, $R_{OUT}[k]$:

$$\begin{bmatrix} L_{OUT}[k] \\ R_{OUT}[k] \end{bmatrix} = \begin{bmatrix} L[k]+\varepsilon C[k] \\ R[k]+\varepsilon C[k] \end{bmatrix}$$

Here, e denotes a weight that determines the strength of $C[k]$ in the downmix (typically 0.707). A parameter IID_C that describes the power of C with respect to the power of L and R is extracted:

$$IID_C = 10\log 10\left(\frac{\varepsilon^2\sum_k C[k]C^*[k]}{\sum_k L[k]L^*[k]+\sum_k R[k]R^*[k]}\right)$$

The process as described above is repeated for each time/frequency tile. Subsequently, the signals $L_{OUT}[k]$ and $R_{OUT}[k]$ are transformed to the time domain and

combined with previous segments using overlap-add, resulting in the output signals $l_{OUT}[n]$ and $r_{OUT}[n]$. A schematic overview of the encoder is shown in Fig. A1.

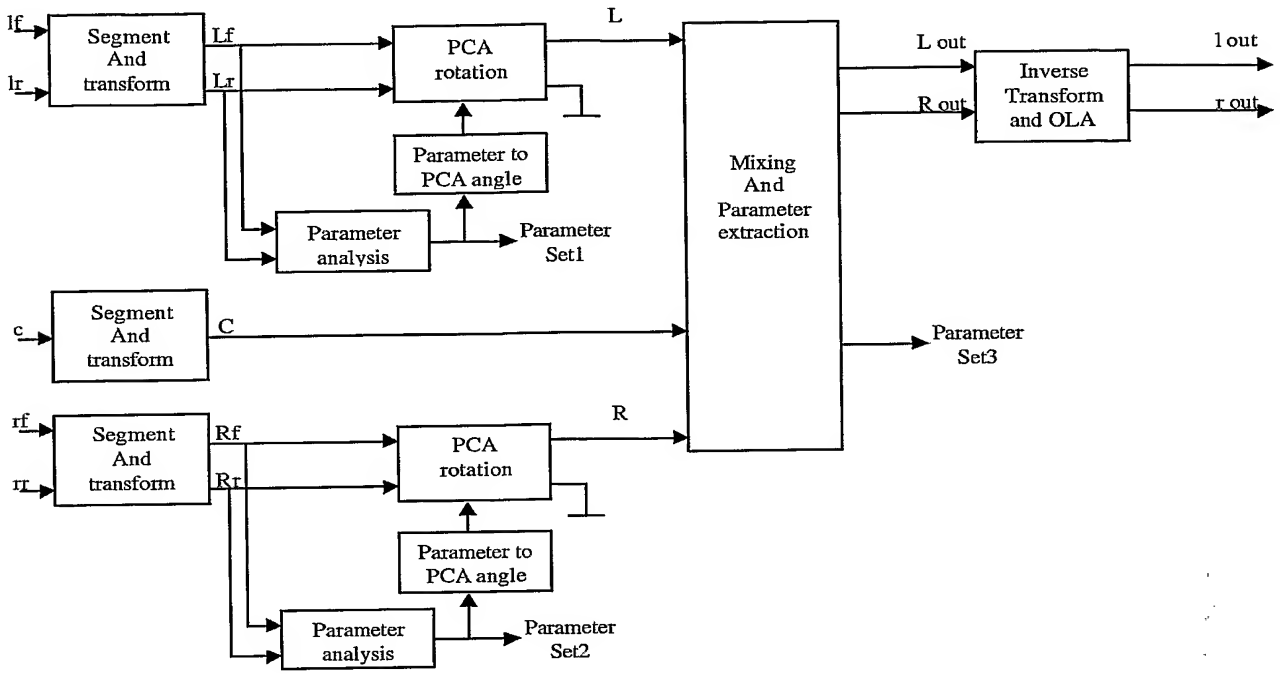

Fig. A1. Schematic overview of the encoder

Decoder for stereo playback

For stereo playback, the transmitted signals $l_{OUT}[n]$ and $r_{OUT}[n]$ are reproduced over the two playback channels, without further processing.

Decoder for 3-channel playback

For 3-channel playback, the two received channels $l_{OUT}[n]$ and $r_{OUT}[n]$ are segmented and transformed to the frequency domain. Subsequently, the output signals $L[k]$, $R[k]$ and $C[k]$ are obtained as follows:

$$\begin{bmatrix} L[k] \\ R[k] \\ C[k] \end{bmatrix} = \begin{bmatrix} w_L L_{OUT} \\ w_R R_{OUT} \\ w_{LC} L_{OUT} + w_{RC} R_{OUT} \end{bmatrix}$$

with

$$w_L = \sqrt{1 - \frac{\sigma_c^2}{\sigma_L^2}}$$

$$w_R = \sqrt{1 - \frac{\sigma_c^2}{\sigma_R^2}}$$

$$w_{LC} = \frac{0.5}{\varepsilon}\sqrt{\frac{\sigma_c^2}{\sigma_L^2}}$$

$$w_{RC} = \frac{0.5}{\varepsilon}\sqrt{\frac{\sigma_c^2}{\sigma_R^2}}$$

$$\sigma_L^2 = \sum_k L[k]L^*[k]$$

$$\sigma_R^2 = \sum_k R[k]R^*[k]$$

$$\sigma_C^2 = \frac{\sigma_L^2 + \sigma_R^2}{2 + 10^{-IID_C/10}}$$

Decoder for 5-channel playback

For five-channel playback, first the 3-channel playback reconstruction as described above is performed, resulting in signals $L[k]$, $R[k]$ and $C[k]$. The next step entails splitting $L[k]$ in $L_f[k]$ and $L_r[k]$, and splitting $R[k]$ in $R_f[k]$ and $R_r[k]$. This splitting process is performed using the inverse PCA rotation as used in the encoder. The dominant $Y[k]$ and residual $Q[k]$ signal, which are required for the inverse PCA rotation, are obtained as follows:

$$\begin{bmatrix} Y[k] \\ Q[k] \end{bmatrix} = \begin{bmatrix} L[k]\cos\gamma \\ H[k]L[k]\sin\gamma \end{bmatrix}$$

Here, $H[k]$ denotes an all-pass decorrelation filter to obtain a decorrelated version of $L[k]$. The angle *?* is given by:

$$\gamma = \arctan\left(\frac{1-\sqrt{\mu}}{1+\sqrt{\mu}}\right)$$

Subsequently, the signals $L_f[k]$ and $L_r[k]$ are obtained using inverse PCA:

$$\begin{bmatrix} L_f[k] \\ L_r[k] \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \exp(jOPD_L) & 0 \\ 0 & \exp(jOPD_L - IPD_L) \end{bmatrix} \begin{bmatrix} Y[k] \\ Q[k] \end{bmatrix}$$

This process is then repeated for the right channel.

Decoder for 4-channel playback

The decoder for 4-channel playback can be simply obtained by first decoding 5 channels (l_f, l_r, c, r_f, and r_s), followed by mixing of the centre channel (c) in front left and front right:

$$l_{f,playback} = l_f + 0.707\ c$$

$$r_{f,playback} = r_f + 0.707\ c$$

The factor 0.707 ensures that the total power of the centre channel is constant, independent of playback through the single centre speaker or as phantom source created by left front and right front. The surround channels remain unchanged (i.e., the signals are the same as for 5-channel playback).

Annex B Coding the low frequency effects (LFE) channel in parametric multichannel audio coding

The multi-channel audio coder described in on the previous pages, does not support the coding of a low frequency effects (LFE) channel, which is incorporated in a standard commonly known as 5.1.

The proposed coder incorporates the LFE channel. To this end additional parameters need to be sent to the decoder. Also, the same time, the 2-channel down-mix is modified to enable the decoding of the LFE channel. This is done in such a way that the good quality of the stereo image of the 2-channel down-mix is preserved.

The parameters are typically analysed as a function of time and frequency (i.e., for a set of time/frequency tiles). The bandwidth of the LFE is typically limited to approximately 120 Hz. The proposed solution allows for a variable bandwidth of the LFE channel.

A multichannel coder is described in Annex A. A schematic overview of its encoder is shown in Fig. 1. in the following, only the changes required for incorporating the LFE channel are explained.

Encoder

In Fig. B2, a schematic overview of the encoder incorporating the subwoofer channel is shown.

Assume that $c[n]$ and $lfe[n]$ describe the discrete time-domain waveforms for the centre and the LFE signal respectively. The signal Cs and the parameters from the block 'parameter analysis' are obtained from signals c and lfe in the same way in which the signal L and the parameters from the block 'parameter analysis' are obtained from signals lf and lr, as described in Annex A. There is however a difference. Because of the low-frequency behaviour of the signal lfe, this is done only for a limited number of frequency subbands. To this end, a parameter describing the number of frequency subbands occupied by the signal lfe is incorporated in parameter set 4. In the remaining higher subbands, only the signal C is transmitted. The other parameters included in parameter set 4 are the level difference (IID) and possibly the phase difference (IPD) between the centre and the LFE channel. If the IPD is sent, also the OPD parameter needs to be transmitted.

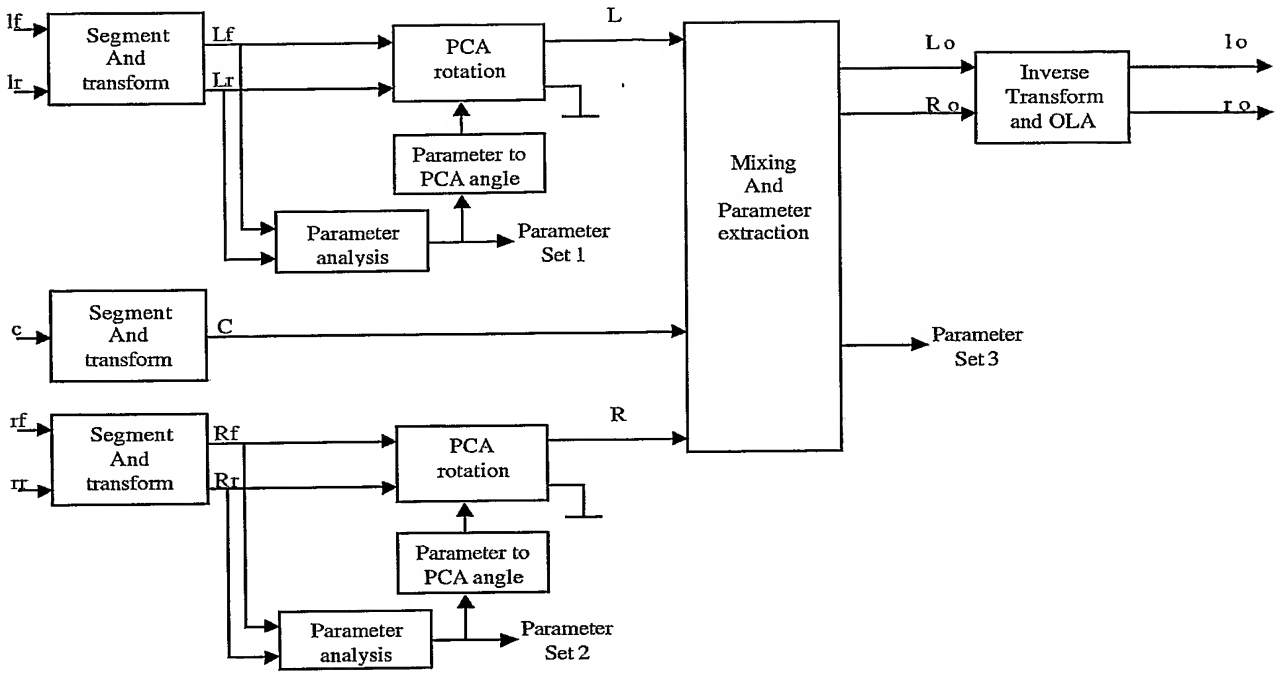

Fig. B1. Schematic overview of the encoder described in Annex A

Decoder for 5.1-channel playback

For 5.1-channel playback, the signals $C[k]$ and $LFE[k]$ are obtained from the reconstructed signal $Cs[k]$ similar to the way in which signals $Lf[k]$ and $Lr[k]$ are obtained from $L[k]$ as described in Annex A. The values of the not transmitted parameter ICC are set to 1. If the IPD and OPD are not transmitted, they are set to 0.

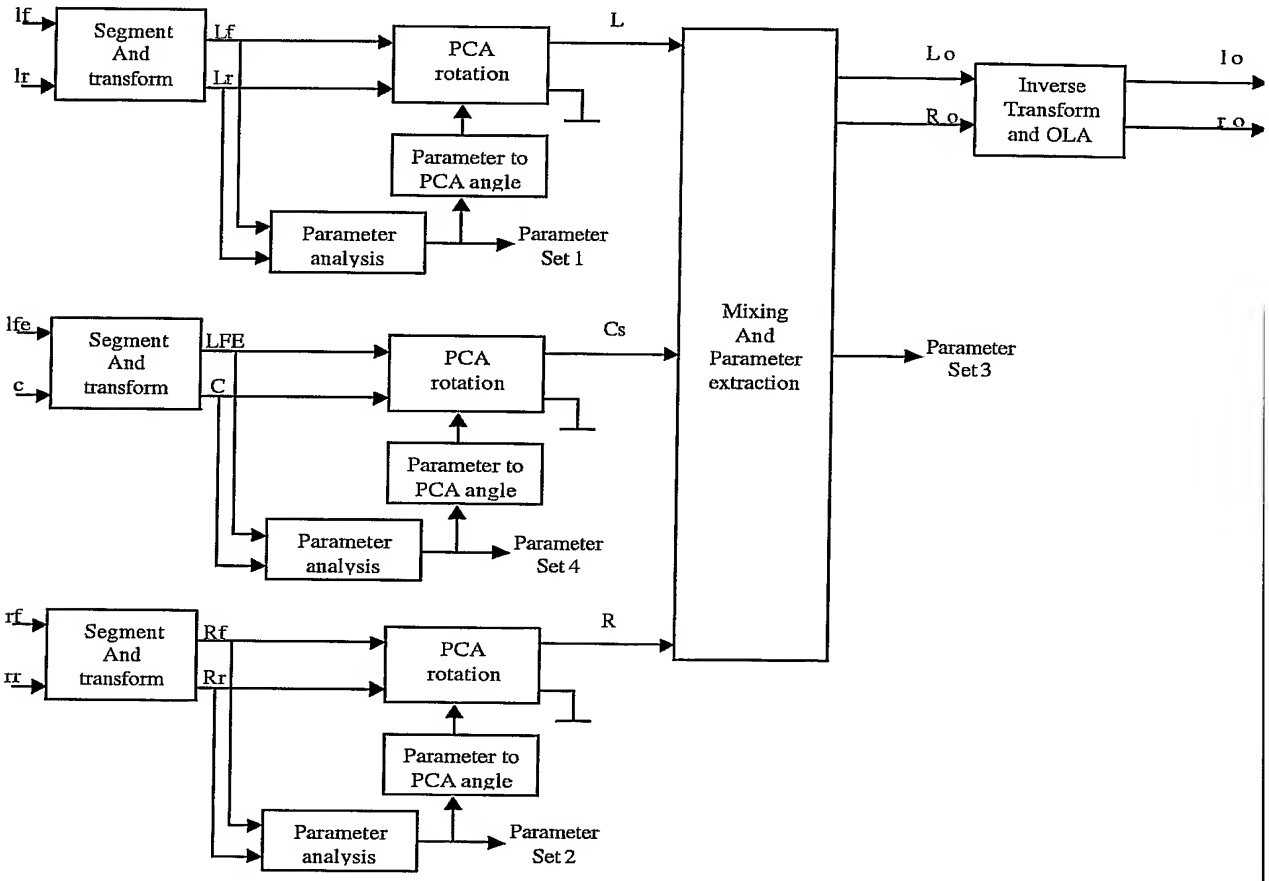

Fig. B2. Schematic overview of the encoder including an LFE channel

Annex C, Reducing the inter-channel interference

The multi-channel audio coder described in Annexes A and B aims at: Approximating the multi-channel audio with two channels and parametric overhead. This is done for low bit rate reasons.

Backwards compatibility with 2, 3, 4 and 5-channel reproduction systems. To this end a good quality of the stereo image of the 2 transmitted channels is required.

The multi-channel audio coder described in Annexes A and B exhibits a significant amount of inter-channel interference, mainly due to the demand for a good stereo image of the 2-channel down-mix. The current invention significantly reduces the inter-channel interference. In this way, the quality of the reproduced multi-channel audio is significantly enhanced.

The coder proposed in this Annex C represents N input channels by 2 down-mix channels and parametric overhead. In order to get the best possible reconstruction, in the sense of least-square-errors, of the N input channels at the decoder using only 2 channels, Principal Component Analysis (PCA) should be used. Perfect reconstruction of the N input channels at the decoder is only possible if all N channels from PCA are used. Drawback of PCA is the fact that no control can be exerted over the 2 down-mix channels. This means that the abovementioned requirement regarding the good quality of the stereo image is not met when employing PCA.

As in the case of PCA, also in the case of employing 2 down-mix channels that do have a good quality of the stereo image, a perfect reconstruction at the decoder is only possible when these 2 down-mix channels are extended with an appropriate set of N-2 channels. As opposed to PCA, whose N channels are orthogonal so that the N-2 discarded channels cannot be predicted using the 2 down-mix channels, now the N-2 channels can – to some extent – be predicted from the 2 down-mix channels. The proposed coder exploits this predictability at the decoder. In order to do so, parameters need to be sent to the decoder. These parameters are typically analysed as a function of time and frequency (i.e., for a set of time/frequency tiles).

As compared to Annex A or Annex B the following changes are made in the proposed coder:

- at the encoder: the parameters required at the decoder have to be computed.
- in the bit-stream: the parameters required at the decoder have to be included.
- at the decoder: a parameter dependent up-mix from 2 to 5 channels is performed (in Annex A, a fixed up-mix is performed).

Encoder

Assume an N-channel audio signal, where $z_1[n]$, $z_2[n]$, ..., $z_N[n]$ describe the discrete time-domain waveforms of the N channels. These N signals are segmented using a common segmentation, preferably using overlapping analysis windows. Subsequently, each segment is converted to the frequency domain using a transform (e.g., FFT). However, filter-bank structures may also be appropriate to obtain time/frequency tiles. This process results in segmented, sub-band representations of the input signals, which will be denoted by $Z_1[k]$, $Z_2[k]$, ... , $Z_N[k]$ with k denoting frequency index.

From these N channels, 2 down-mix channels are created, being $L_0[k]$ and $R_0[k]$, which are also segmented, sub-band representations. Each down-mix channel is a linear combination of the N input signals:

$$L_0[k] = \sum_{i=1}^{N} \alpha_i Z_i[k],$$

$$R_0[k] = \sum_{i=1}^{N} \beta_i Z_i[k].$$

The parameters α_i and β_i can be set on the basis of a certain criterion. In ID 695741, this criterion is the good stereo image of the stereo signal consisting of $L_0[k]$ and $R_0[k]$.

Perfect reconstruction of the N input channels at the decoder is only possible when these 2 down-mix channels are extended with an appropriate set of N-2 channels. As opposed to PCA, whose N channels are orthogonal so that the N-2 discarded channels cannot be predicted using the 2 most relevant channels, now the N-2 channels can – to some extent – be predicted from the 2 down-mix channels.

If the N-2 discarded channels are denoted by $C_{0,i}[k]$, then these channels are predicted from the two down-mix channels by:

$$\hat{C}_{0,i}[k] = \tilde{C}_{1,i} L_0[k] + \tilde{C}_{2,i} R_0[k].$$

For choosing parameters $\tilde{C}_{1,i}$ and $\tilde{C}_{2,i}$ various optimisation criteria are possible. We choose as an optimisation criterion the minimal Euclidian norm of the difference of signal $C_{0,i}[k]$ and its estimation $\hat{C}_{0,i}[k]$. Parameters $\tilde{C}_{1,i}$ and $\tilde{C}_{2,i}$ need to be sent to the decoder.

It can be shown that the parameters $\tilde{C}_{1,i}$ and $\tilde{C}_{2,i}$ are related to the parameters that are obtained when minimising the Euclidian norm of the difference of the original input

channel $Z_i[k]$ and its estimation at the decoder $\hat{Z}_i[k]$. A coder that uses these latter parameters is further described.

The square of the Euclidian norm of the difference of the original input channel $Z_i[k]$ and its estimation at the decoder $\hat{Z}_i[k]$ can be written as:

$$\sum_k \left| Z_i[k] - \hat{Z}_i[k] \right|^2,$$

with

$$\hat{Z}_i[k] = C_{1,Z_i} L_0[k] + C_{2,Z_i} R_0[k].$$

Minimisation of $\sum_k \left| Z_i[k] - \hat{Z}_i[k] \right|^2$ leads to the following expressions:

$$C_{1,Z_i} = \frac{< L_0[k], Z_i[k] >^* \left\| R_0[k] \right\|^2 - < R_0[k], Z_i[k] >^* < L_0[k], R_0[k] >^*}{\left\| L_0[k] \right\|^2 \left\| R_0[k] \right\|^2 - \left| < L_0[k], R_0[k] > \right|^2},$$

$$C_{2,Z_i} = \frac{< R_0[k], Z_i[k] >^* \left\| L_0[k] \right\| - < L_0[k], Z_i[k] >^* < L_0[k], R_0[k] >}{\left\| L_0[k] \right\|^2 \left\| R_0[k] \right\|^2 - \left| < L_0[k], R_0[k] > \right|^2},$$

with

$$\left\| A[k] \right\|^2 = \sum_k \left| A[k] \right|^2,$$

$$< A[k], B[k] >= \sum_k A[k] B^*[k].$$

For the coefficients $C_{1,Zi}$ and $C_{2,Zi}$, the following relations can be derived:

$$\sum_{i=1}^{N} \alpha_i C_{1,Z_i} = 1,$$

$$\sum_{i=1}^{N} \beta_i C_{2,Z_i} = 1,$$

$$-\sum_{i=1}^{N} \beta_i C_{1,Z_i} = 0,$$

$$-\sum_{i=1}^{N} \alpha_i C_{2,Z_i} = 0.$$

Having N channels, with 2 parameters per channel ($C_{1,Zi}$ and $C_{2,Zi}$ for the i-th channel), but at the same time 4 equations describing relations between these parameters, $2N$-4 parameters need to be sent to the decoder.

The process described above is repeated for each time/frequency tile. Subsequently, the signals $L_0[k]$ and $R_0[k]$ are transformed to the time domain and combined with previous segments using overlap-add, resulting in the output signals $l_0[n]$ and $r_0[n]$.

Summarising, the encoder sends 2 down-mix channels, $l_0[n]$ and $r_0[n]$, and for each time/frequency tile $2N$-4 parameters, that describe how to retrieve the input channels from the 2 down-mix channels, to the decoder.

Decoder

At the decoder side, for each time/frequency tile, first the coefficients $C_{1,Zi}$ and $C_{2,Zi}$ are computed for all N channels, using the $2N$-4 coefficients that are transmitted and the 4 equations describing relations between the coefficients. Then each input channel $Z_i[k]$ is approximated by $\hat{Z}_i[k]$, with

$$\hat{Z}_i[k] = C_{1,Z_i} L_0[k] + C_{2,Z_i} R_0[k],$$

where $L_0[k]$ and $R_0[k]$ are the received 2 down-mix channels.

Incorporation of the coder in the multi-channel coder described in Annex A or Annex B

A schematic overview of the encoder of the multi-channel coder described in Annex A and Annex B is given in Fig. A1 and Fig. B2 respectively. The coder described in this Annex C, can be used to replace the block called "Mixing And Parameter extraction", that has as inputs the channels L, R and Cs and as outputs the channels L_0 and R_0 and parameter set 3. In order to get a good stereo image of the 2 down-mix channels L_0 and R_0, they are chosen as:

$$L_0[k] = L[k] + Cs[k],$$
$$R_0[k] = R[k] + Cs[k].$$

Because of the three input channels (hence $N = 3$), only $2N$-4, or 2 parameters need to be transmitted to the decoder. It is advantageous to transmit 2 parameters that have the same range (e.g. $C_{1,L}$ and $C_{2,R}$), so that the same quantisation can be applied to them.

At the decoder side, for 3 or more channel playback, first all 6 parameters ($C_{1,L}$, $C_{2,L}$, $C_{1,R}$, $C_{2,R}$ $C_{1,Cs}$ and $C_{2,Cs}$) are computed using the 2 transmitted parameters and the relations between the 6 parameters. For example, if $C_{1,L}$ and $C_{2,R}$ are transmitted, then it

follows that $C_{2,L} = C_{2,R}$ -1, $C_{1,R} = C_{1,L}$–1, $C_{1,Cs} = 1 - C_{1,L}$ and $C_{2,Cs} = 1 - C_{2,R}$. The output signals $\hat{L}[k]$, $\hat{R}[k]$ and $\hat{C}s[k]$ are obtained as follows:

$$\begin{bmatrix} \hat{L}[k] \\ \hat{R}[k] \\ \hat{C}s[k] \end{bmatrix} = \begin{bmatrix} C_{1,L}L_0[k] + C_{2,L}R_0[k] \\ C_{1,R}L_0[k] + C_{2,R}R_0[k] \\ C_{1,C}L_0[k] + C_{2,C}R_0[k] \end{bmatrix}.$$

Playback of 4- or 5-channels is explained in Annex A.

Annex D Improved stereo coding

Traditional coding schemes, like e.g. MPEG-1 Layer III (mp3, [1]) employ stereo coding tools to improve the coding efficiency. One of these coding tools is known as Mid/Side (M/S) stereo coding or Sum/Difference stereo coding [2]. Using M/S coding a stereo signal consisting of a left signal $l[n]$ and a right signal $r[n]$ is coded as a sum signal $m[n]$ and a difference signal $s[n]$[1]:

$$m[n] = r[n] + l[n]$$
$$s[n] = r[n] - l[n]$$

For (almost) identical signals $l[n]$ and $r[n]$ this gives a large coding gain as the corresponding difference signal $s[n]$ is close to being zero, whereas the sum signal contains practically all the signal energy. Hence, in this situation the bit rate required for coding the sum and difference signals is close to the bit rate required for coding only a single channel.

Alternatively the mid-side coding process can be described by means of a rotation matrix:

$$\begin{pmatrix} m[n] \\ s[n] \end{pmatrix} = c \begin{pmatrix} \cos\left(\frac{\pi}{4}\right) & \sin\left(\frac{\pi}{4}\right) \\ -\sin\left(\frac{\pi}{4}\right) & \cos\left(\frac{\pi}{4}\right) \end{pmatrix} \begin{pmatrix} l[n] \\ r[n] \end{pmatrix}$$

It is obvious that the left and right signals have been rotated over an angle of $\pi/4$. This is illustrated in Figure D1. The sum signal can be interpreted as a projection of the left and right samples onto the line l=r, whereas the difference signal can be interpreted as a projection of the left and right samples onto the line l=-r.

[1] Usually the sum and difference are calculated as $m[n] = c \cdot (l[n] + r[n])$ and $s[n] = c \cdot (l[n] - r[n])$. For explanatory reasons, the constant c has been discarded and the sign of $s[n]$ has been inverted.

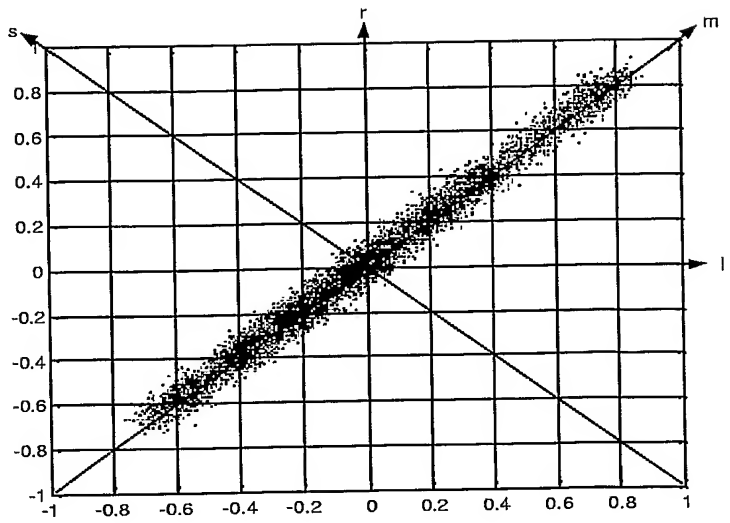

Figure D1: Rotation of the left and right signals $l[n]$ and $r[n]$ over an angle of $\pi/4$.

In order to obtain a minimum signal power in the residual signal (i.e., maximum coding gain) for a wide class of input signals, the rotation angle needs to be variable. An improved signal mapping, applied in a sub-band coding system using a variable rotation angle, is outlined in [3, 4]. The following signal mapping is applied:

$$\begin{pmatrix} m'[n] \\ s'[n] \end{pmatrix} = c \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} l[n] \\ r[n] \end{pmatrix},$$

where $m'[n]$ and $s'[n]$ represent the dominant and the residual signal respectively and the angle α is chosen to minimize the power of the signal $s'[n]$. Due to the unitary rotation operation, the power of m'[n] is then maximized. This process is illustrated in Figure D2.

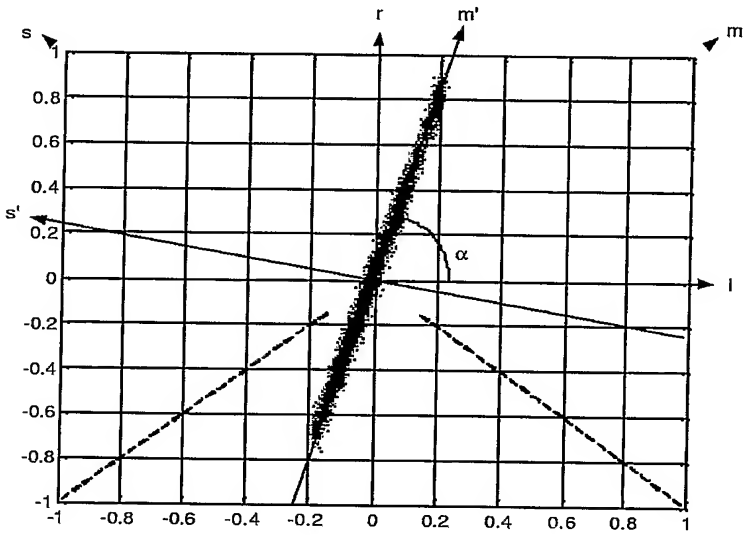

Figure D2: Left and right signal mapping using a rotation along an angle α

Using regular M/S coding (as represented by m and s in Figure D1), the signal illustrated in Figure D2 would still result in a residual signal with considerable energy. As such, the coding gain obtained by M/S coding would be marginal. However, using the variable rotation angle α as illustrated above, a very small residual signal can be obtained. Obviously, this rotation technique works particularly well when the left signal is approximately a scaled version of the right signal.

Both the M/S coding technique (i.e., rotation with a fixed angle) as well as the variable rotation technique described above are typically not applied to the broadband signal, but rather to signals (or frequency domain representations) representing only a smaller part of the full bandwidth of the audio signal, as e.g. described in [3, 4].

Although the rotation technique as described in [3,4] eliminates much of the disadvantages of M/S coding it is still sub optimal for signals having a strong phase or time offset. This is illustrated in Figure D3.

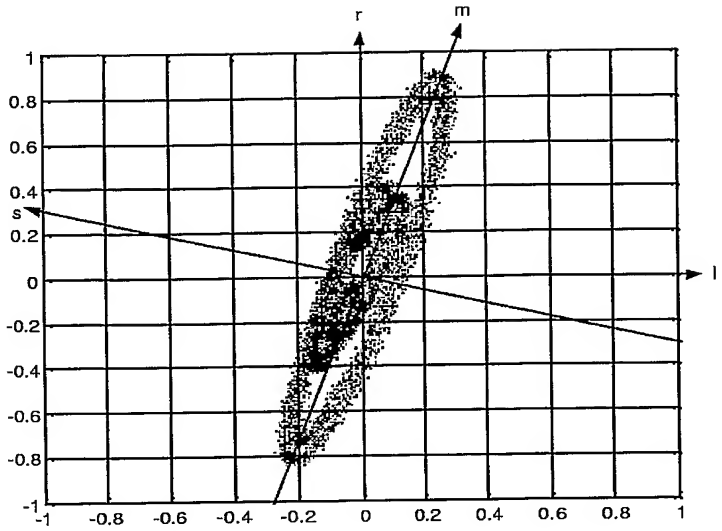

Figure D3: Rotation of left and right signals along an angle α.

The oval-like structure indicates a time or phase delay between l and r. Because the rotation implies a real-valued projection, the residual signal will still contain a significant amount of energy (although usually less than using regular M/S coding).

In order to further reduce the residual signal energy it is proposed to extend the current signal rotation by employing complex-valued phase rotations to the left and right signal components. From this point, it is assumed that the left and right signals are represented by their complex-valued frequency domain representations $l[k]$ and $r[k]$. One method to obtain such signal representation is as follows. First the left and right time domain signal segments are windowed:

$$l_q[n] = l[n+qH] \cdot h[n]$$
$$r_q[n] = r[n+qH] \cdot h[n]$$

where q represents the frame index ($q = 0,1,2,....$), H represents the hop-size or update-size and $n = 0...L-1$ where L equals the length of window $h[n]$. These windowed segments are then transformed to the frequency domain by means of a Discrete Fourier Transform (DFT):

$$l[k] = \sum_{n=0}^{N-1} l_q[n] \cdot \exp\left(-j\frac{2\pi kn}{N}\right)$$
$$r[k] = \sum_{n=0}^{N-1} r_q[n] \cdot \exp\left(-j\frac{2\pi kn}{N}\right),$$

where N represents the DFT length ($N \geq L$). Because of symmetry in the DFT only the first $N/2+1$ points are preserved. Furthermore, in order to obtain energy preservation, the first DFT points are scaled:

$$l[0] \equiv l[0]/2$$
$$r[0] \equiv r[0]/2$$

The signal model is extended in the following way:

$$\begin{pmatrix} m''[k] \\ s''[k] \end{pmatrix} = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} e^{j\varphi_1} & 0 \\ 0 & e^{j(\varphi_1-\varphi_2)} \end{pmatrix} \begin{pmatrix} l[k] \\ r[k] \end{pmatrix}.$$

As can be observed from the equation above, the real-valued rotation (using a variable rotation angle a is extended with a complex-valued phase modification matrix. The angle φ_2 is used to minimize the energy of the residual signal by (phase-) rotating the right signal. The common angle φ_1 can be used to maximize the continuation of the signal over frame boundaries.

After signal mapping/modification the dominant and residual time domain signals $m[n]$ and $s[n]$ are obtained by first applying the inverse DFT on the frequency domain representations $m[k]$ and $s[k]$:

$$m_q[k] = \sum_{n=0}^{N-1} m[k] \cdot \exp\left(j\frac{2\pi kn}{N}\right)$$
$$s_q[k] = \sum_{n=0}^{N-1} s[n] \cdot \exp\left(j\frac{2\pi kn}{N}\right)$$

where the dominant and residual frequency domain representations $m[k]$ and $s[k]$ have been zero-padded to length N. The time domain signals are then obtained by means of overlap-add:

$$m[n+qH] \equiv m[n+qH] + 2\Re\{m_q[n] \cdot h[n]\}$$
$$s[n+qH] \equiv s[n+qH] + 2\Re\{s_q[n] \cdot h[n]\}$$

Alternatively, complex-modulated filter banks could be employed to obtain a complex-valued frequency domain representation.

As an example, the following synthetic signal is mapped by the three different methods described above:

$$l[n] = 0.5\cos(0.32n + 0.4) + 0.05 \cdot z_1[n] + 0.06 z_2[n]$$
$$r[n] = 0.25\cos(0.32n + 1.8) + 0.03 \cdot z_1[n] + 0.05 z_3[n]$$

where $z_1[n]$, $z_2[n]$ and $z_3[n]$ are independent white noise sequences with unit variance. Part of the signals $l[n]$ and $r[n]$ are shown in Figure D4.

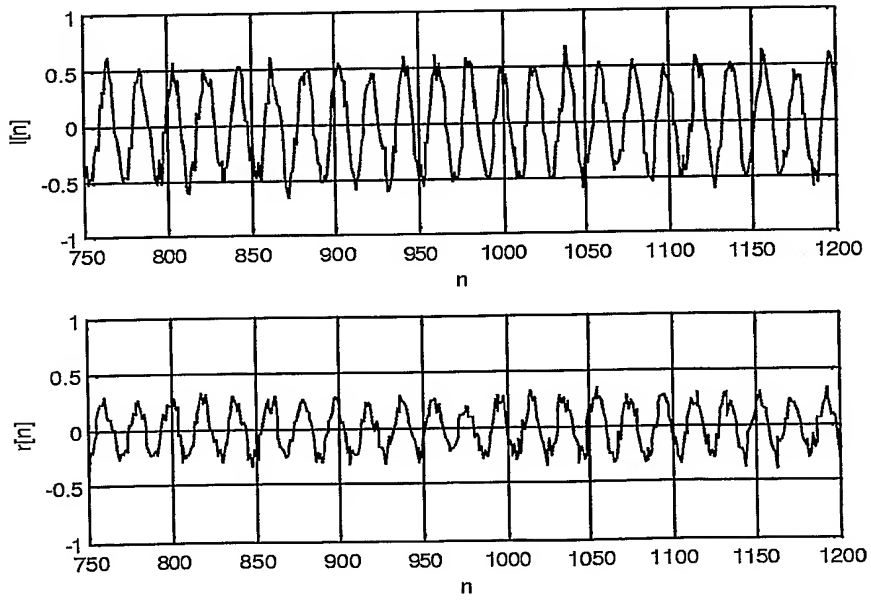

Figure D4: Left and right signals l[n] and r[n] respectively

Figure D5, Figure D6 and Figure D7 show the results for the M/S transform, the signal rotation over an (optimal) angle α and the rotation over both an (optimal) angle α *and* phase rotation as proposed in this ID respectively. In this particular example the angles α, φ_1 and φ_2 are fixed. In a typical embodiment, these parameters are both time and

frequency dependent. In each figure, the top panel represents the dominant signal (m[n]); the bottom panel shows the residual signal (s[n]).

The M/S mapping, as shown in Figure D5, clearly does not increase the coding efficiency for this particular situation. As a matter of fact, the residual signal energy, i.e., the energy of the signal $s[n]$, is higher than the energy of the input signal r[n].

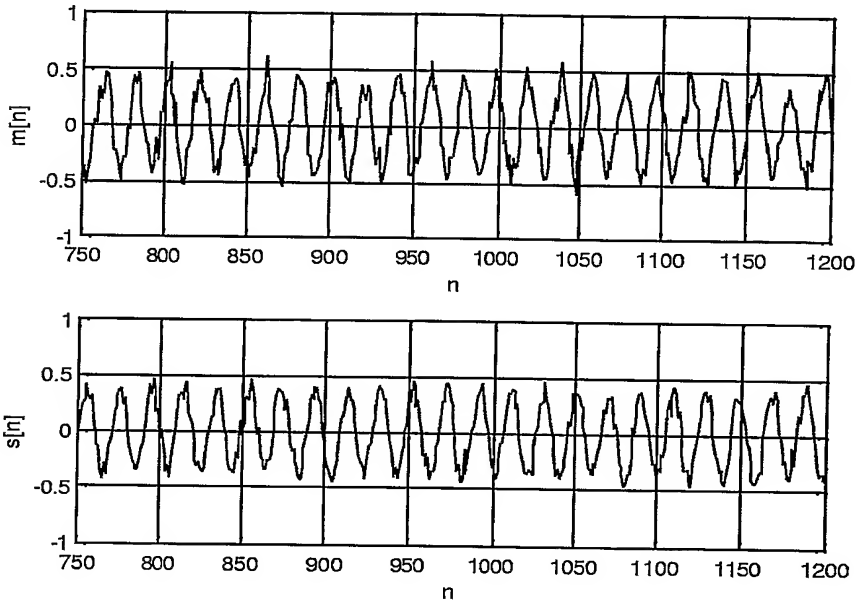

Figure D5: Resulting signals after M/S mapping

The mapping by means of applying the (optimal) signal rotation α as illustrated in Figure D6 does also not help for this particular signal. Only a negligible energy reduction is obtained.

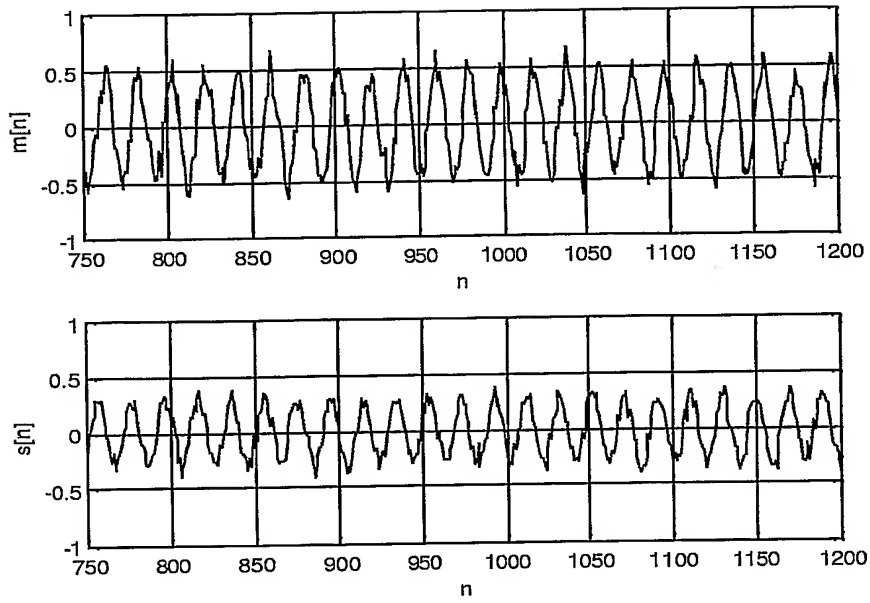

Figure D6: Resulting signals after rotation

Finally, the results of the extension of the mapping as proposed in this ID are shown in Figure D7. Here a clear reduction of residual signal energy is observed.

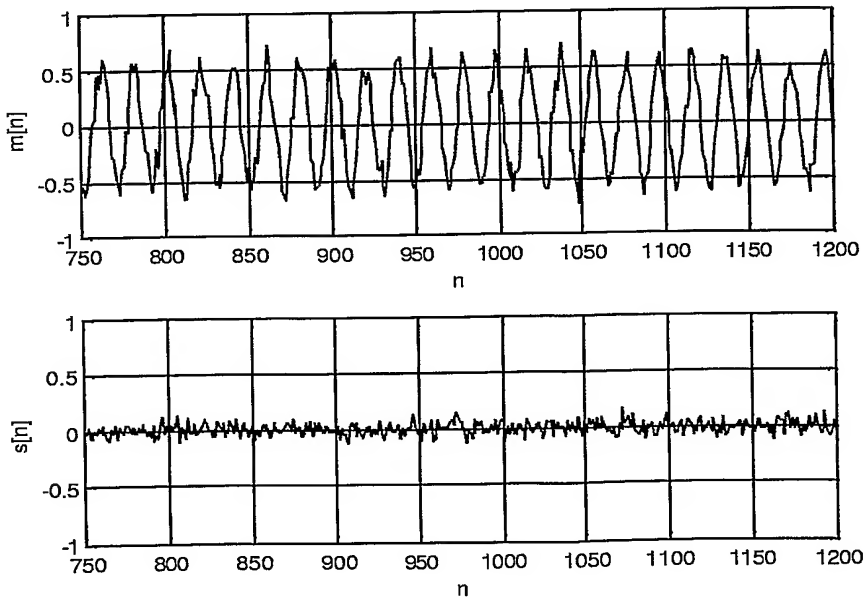

Figure D7: Resulting signals after signal *and* phase rotation

A block diagram of an encoder according to the invention is given in Figure D8. The left and right frequency domain representations l and r are phase rotated to obtain

a maximum coherence (angle φ_2) and

an optimal signal continuation over time (angle φ_1).

Consequently, the phase-rotated left and right signals are rotated over an angle α for maximal reduction of the residual signal error as described above. The parameters α, φ_1 and φ_2 are quantized and coded into the bit stream. The dominant signal *m* and the residual signal *s* can be coded by two independent conventional mono audio coders (or of course one dual mono encoder). Additionally, certain parts of the time-frequency plane of the signal *s*, not perceptually contributing to the final output signal, can be discarded in the time-frequency (t/f) selector unit. The overall bit stream is formed by merging the bit stream corresponding to the dominant signal *m*, the residual signal *s* and the parameter bit stream.

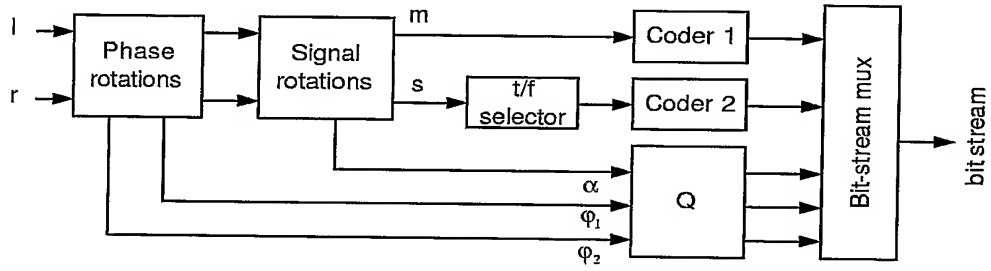

Figure D8: Block diagram of proposed encoder

Figure D9 shows a block diagram of the decoder corresponding to the encoder of Figure D8. First the bit stream is de-multiplexed into separate bit streams for the dominant signal, the residual signal and the parameters. The bit streams for the dominant signal and the residual signal are decoded resulting in the signals *m*' and *s*'. Then the inverse rotation ($-\alpha$) is applied to obtain preliminary left and right signal representations. Finally the left and right signals, *l*' and *r*' respectively, are obtained by applying the inverse phase rotations ($-\varphi_1$ and $-\varphi_2$).

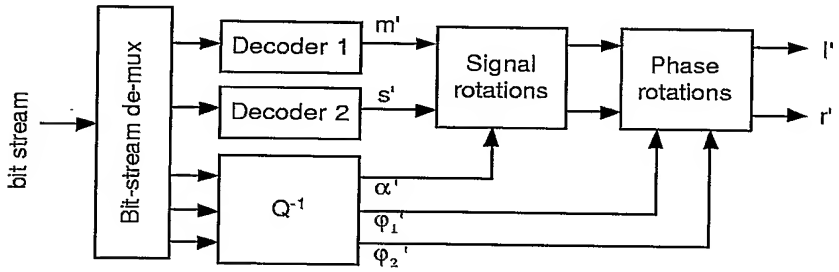

Figure D9: Block diagram of proposed decoder

The invention can also be advantageously applied in combination with a parametric stereo coding system [5]. A general block diagram of a parametric stereo coding scheme is given in Figure D10. It is fairly similar to the block diagram of the proposed encoder except for the fact that:

- no residual signal is being transmitted and
- the angle α is not transmitted, but instead an IID value and a coherence value ? are transmitted.

The IID value represents the Inter-channel Intensity Difference, denoting the (frequency and time variant) intensity differences between the left and right input channels. The coherence value denotes the coherence, i.e., the similarity, between the left and right input channels after phase synchronization. The angle α can be derived from the IID and coherence value.

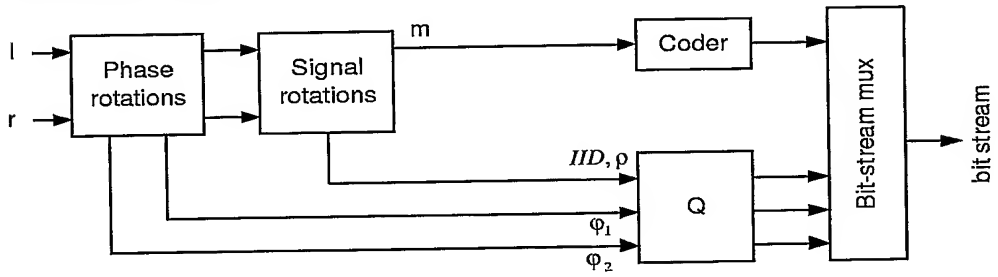

Figure D10: Block diagram of parametric stereo encoder

A corresponding decoder block diagram is shown in Figure D10. It corresponds to the block diagram of Figure D9 except for:

- the residual signal *s'* is now estimated based on the dominant signal *m'* by means of a de-correlation process D and
- the amount of coherence between the left and right output signals is determined by a scaling operation.

The scaling operation basically describes the ratio between the dominant signal and the residual signal.

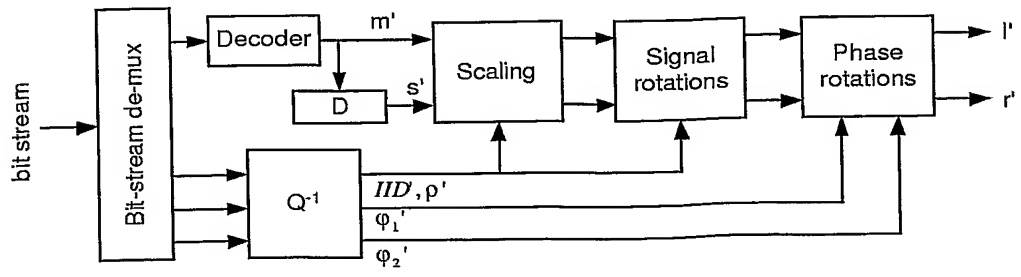

Figure D11: Block diagram of parametric stereo decoder

Figure D12 shows a block diagram of the parametric stereo encoder enhanced with residual coding. With respect to Figure D10, the only difference resides in the coding of (part of) the residual signal *s*. Which part of the residual signal is coded by Coder 2, is determined by the time-frequency (t/f) selector unit.

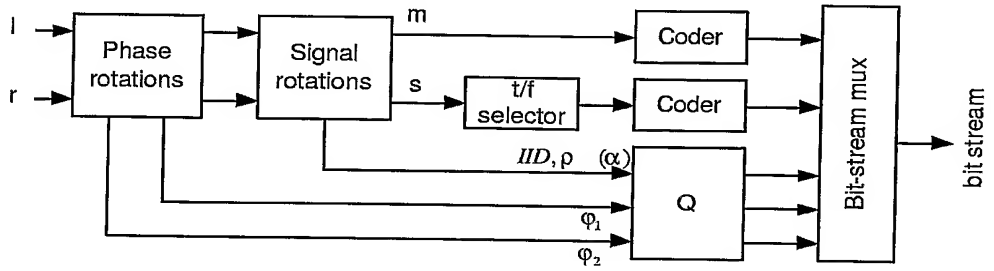

Figure D12: Block diagram of enhanced parametric stereo encoder

A block diagram of the parametric stereo decoder enhanced with residual coding is shown in Figure D13. The bit stream is first de-multiplexed into separate streams for the dominant signal, the residual signal and the stereo parameters. The dominant and residual signals are then decoded by Decoder 1 and Decoder 2, respectively. Those spectral/temporal parts of the residual signal which have been coded are signalled either:
implicitly, by detecting "empty" areas in the time-frequency plane or
explicitly, by means of bits in the parameter bit stream.

This information is applied in the de-correlation unit D and the Combine unit to fill the empty time-frequency areas in the decoded residual signal with a synthetic residual signal. This synthetic residual signal is generated by using the decoded dominant signal m' and the de-correlation unit D. For all other time-frequency areas, the (transmitted) residual signal is applied to construct the signal *s'*. Note that for these areas, no scaling is applied. Hence, for these areas it can be advantageous to transmit the angle α in the encoder instead of the IID and coherence values as the bit rate required for the single parameter α is smaller than the bit rate required for the IID and coherence parameters. However, transmission of α instead of IID and coherence values makes the system non-backwards compatible to the regular PS system. The subsequent stages of the decoder operate in the same fashion as the conventional parametric stereo decoder.

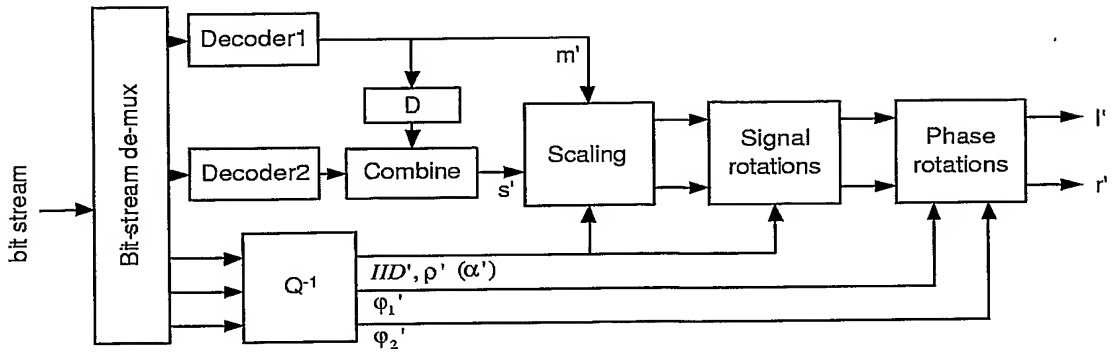

Figure D13: Block diagram of enhanced parametric stereo decoder

The criteria for deciding which time-frequency areas of the residual signal need to be coded are perceptually motivated. If (a time-frequency area of) the residual signal s does not contribute to the audio quality of the final decoder output signal, or if (a time-frequency area of) the de-correlated signal forms a perceptually valid representation of the (corresponding time-frequency area of the) residual signal, it is not necessary to code the residual signal.

By coding different time-frequency aspects of the residual signal in the encoder and by multiplexing the corresponding data into a scalable bit stream, the enhanced parametric stereo codec can be extended to a bit-rate scalable codec. In a scalable system where the layers in the bit stream are dependent, the coded data corresponding to the perceptually most relevant time-frequency aspects should be placed in the base layer, and the less important data moved to refinement, or enhancement, layers. In this case the base layer would consist of the dominant bit stream, a first enhancement layer would consist of the stereo parameters and a second enhancement layer would consist of the residual bit stream.

When layers are removed from the scalable bit stream, and information regarding the residual signal is thus lost, the enhanced parametric stereo decoder can combine the decoded residual signal reconstructed from the data in the remaining layers with the synthetic residual signal in the manner described above to form a meaningful residual signal. Furthermore, if a decoder is not equipped with a second waveform decoder (for the residual signal), e.g. due to complexity restrictions, the signal could still be decoded, although this would result in a lower quality level.

Further bit-rate reductions can be obtained by discarding the values for φ_1 and φ_2 in the bit stream. In that case, the decoder reconstructs the output signals l' and r' using a phase rotation of zero. This method effectively exploits the lack of sensitivity of the human auditory system to high-frequency (inter-aural) phase information.

References

[1] ISO/IEC JTC1/SC29/WG11 MPEG, IS11172-3, Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s, Part 3: Audio, MPEG-1, 1992

[2] J.D. Johnston and A.J. Ferreira, "Sum-difference stereo transform coding," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, San Francisco, CA, March 1992, pp. II:569--572.

[3] R.G. van der Waal and R.N.J Veldhuis, "Subband coding of stereophonic digital signals," in *Proceedings of Int. Conf. Acoustic, Speech, Signal Processing*, pp. 3601-3604, IEEE, 1991.

[4] R.N.J. Veldhuis, R.G. van der Waal, L.M. van de Kerkhof, "Subband coding of a digitial signal in a stereo intensity mode," *US Patent 5,621,855*, Apr. 15, 1997.

[5] J. Breebaart, S. van de Par, A. Kohlrausch and E. Schuijers, "High-quality parametric spatial audio coding at low bitrates," in Proc. 116th AES Convention, Berlin (Germany), May 8-11, 2004.

Annex E Post-processing of the individual multi-channel signals in the stereo down-mix

Many of the multi-channel audio coders as described herein before generate a 2-channel down-mix to be compatible with 2-channel reproduction systems. When this down-mix is created the spatial impression of the original multi-channel mix is lost. The current invention makes improvement of the spatial image of the down-mix possible after creation, based upon the parameters as determined in the multi-channel encoder. Also other post-processing techniques on the individual multi-channel contributions are made possible.

The proposed method makes a reconstruction possible of the multi-channel mix that is not affected by the post-processing. Also post-processing in the decoder is possible for stereo playback as a user-selectable, without the necessity to determine the multi-channel signal first.

Without post-processing the down-mix is comparable with the standard ITU down-mix. The proposed method however improves the down-mix significantly. This is a very important issue, because it is very probable that the quality of the down-mix will be one of the selection criteria within MPEG.

The proposed method is able to determine the contribution in the down-mix of the original channels in the multi-channel mix with the help of the determined parameters in the encoder. In this way post-processing can be applied to specific channels of the multi-channel mix (for example: stereo-widening of the rear channels), whilst the other channels are not affected. The post-processing does not affect the final multi-channel reconstruction. It can also be applied for an improved stereo playback without the necessity to determine the multi-channel mix first.

This method differs from existing post-processing techniques in that it uses the knowledge of the original multi-channel mix (the determined parameters).

Encoder

Assume an N-channel audio signal, where $z_1[n]$, $z_2[n]$,....., $z_N[n]$ describe the discrete time-domain waveforms of the N channels. These N signals are segmented using a common segmentation, preferably using overlapping analysis windows. Subsequently, each segment is converted to the frequency domain using a complex transform (e.g., FFT). However, complex filter-bank structures may also be appropriate to obtain time/frequency

tiles. This process results in segmented, subband representations of the input signals (which will be denoted by $Z_1[k]$, $Z_2[k]$,...., $Z_N[k]$ with k denoting the frequency index).

From these N channels, 2 down-mix channels are created, being $L_O[k]$ and $R_O[k]$. Each down-mix channel is a linear combination of the N input signals:

$$L_O[k] = \sum_{i=1}^{N} \alpha_i Z_i[k]$$

$$R_O[k] = \sum_{i=1}^{N} \beta_i Z_i[k].$$

The parameters α_i and β_i are chosen such that the stereo signal consisting of $L_O[k]$ and $R_O[k]$ has a good stereo image.

In the decoder the N input channels are reconstructed as follows:

$$\hat{Z}_i[k] = C_{1,Z_i} L_O[k] + C_{2,Z_i} R_O[k],$$

where $\hat{Z}_i[k]$ is an estimate of $Z_i[k]$. The parameters C_{1,Z_i} and C_{2,Z_i} are determined in the encoder and transmitted to the decoder.

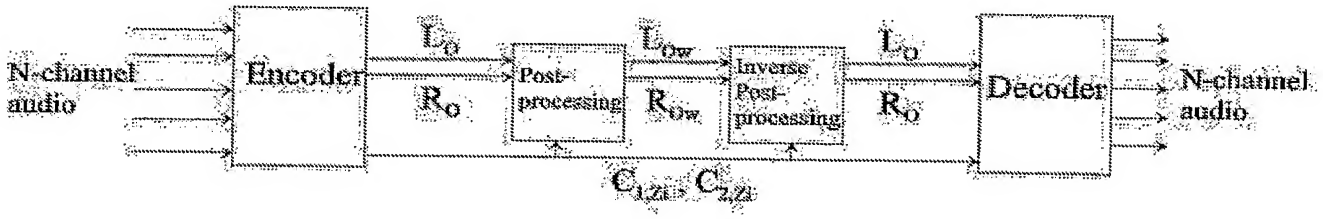

Figure E1: The positioning of the post-processing and inverse post-processing block in the multi-channel coder.

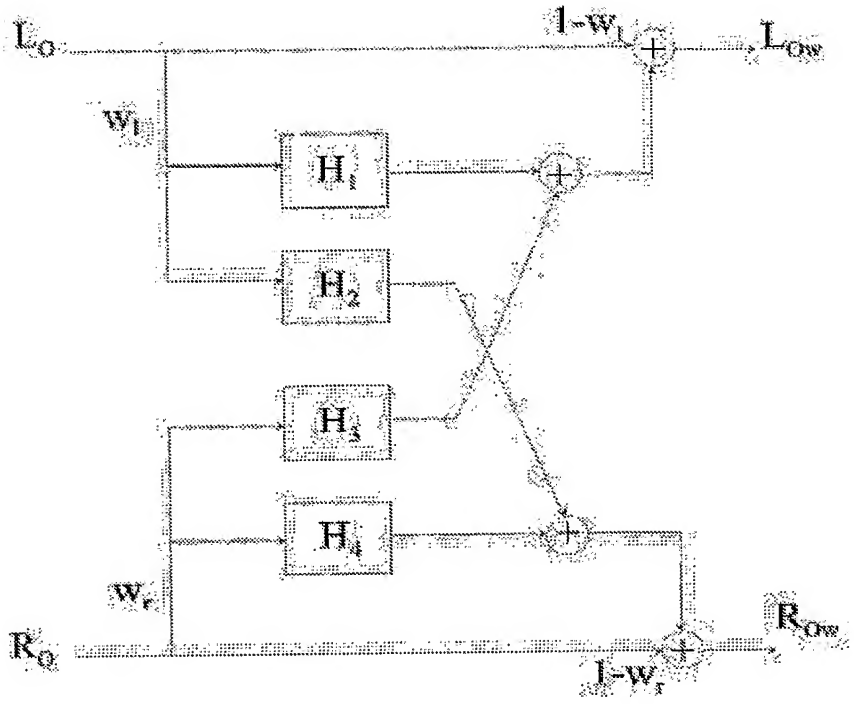

Figure E2: Basic scheme for post-processing of the stereo down-mix.

On the resulting stereo signal post-processing can be applied in a way that it mainly affects the contribution of $Z_i[k]$ in the stereo mix. In Figure E1 the position of this block in the codec is shown.

Figure E2 shows how this post-processing block will look. The parameter w_l determines the amount of post-processing of $L_O[k]$ and w_r of $R_O[k]$. When w_l is equal to 0, $L_O[k]$ is unaffected, and when w_l is equal to 1, $L_O[k]$ is maximally affected. The same holds for w_r with respect to $R_O[k]$.

The following equations hold for the post-processing parameters w_l and w_r:

$$w_l = f_l(C_{1,Z_1},....,C_{1,Z_N},C_{2,Z_1},....,C_{2,Z_N})$$
$$w_r = f_r(C_{1,Z_1},....,C_{1,Z_N},C_{2,Z_1},....,C_{2,Z_N}).$$

The blocks $H_1.....H_4$ in Figure E1 are filters, which can be various types of filters, for example stereo widening filters (as shown in the end of this section). The resulting outputs are:

$$\begin{bmatrix} L_{Ow} \\ R_{Ow} \end{bmatrix} = H \begin{bmatrix} L_O \\ R_O \end{bmatrix},$$

with:

$$H = \begin{bmatrix} 1 - w_l + w_l H_1 & w_r H_3 \\ w_l H_2 & 1 - w_r + w_r H_4 \end{bmatrix}.$$

When the filters $H_1 \ldots\ldots H_4$ are chosen properly, then the matrix H is invertible, when the filters are known in the decoder, because the parameters w_l and w_r can be calculated from the transmitted parameters. So the original stereo signal will be available again which is necessary for decoding of the multi-channel mix.

Another possibility is to transmit the original stereo signal and apply the post-processing in the decoder to make improved stereo playback possible without the necessity to determine the multi-channel mix first.

Incorporation of the coder in a multi-channel coder described in Annex A, B or C

On these pages an encoder is described that codes 5-channel audio. The following equations are applied:

$$L_O[k] = L[k] + Cs[k]$$
$$R_O[k] = R[k] + Cs[k],$$

in which Cs[k] is the mono signal that results after applying OCS between the LFE- (subwoofer) and center-channel. For $L[k]$ and $R[k]$ the following equations hold:

$$L[k] = C_l(\cos(\alpha_l)L_f + \sin(\alpha_l)e^{jIPD_l}L_s)$$
$$R[k] = C_r(\cos(\alpha_r)R_f + \sin(\alpha_r)e^{jIPD_r}R_s),$$

where L_f is the left/front, L_s the left/surround, R_f the right/front and R_s the right/surround channel. The parameters IPD_l and IPD_r (inter-channel phase differences) and C_l and

C_r (complex scaling parameters) are parameters that are determined in the OCS encoder. The angles in these equations can be calculated from the inter-channel intensity differences (IID) and the normalized cross-correlation (IC):

$$\alpha_l = 0.5 \arctan\left(\frac{2IC_l 10^{IID_l/20}}{IC_l^{IID_l/10} - 1}\right)$$
$$\alpha_r = 0.5 \arctan\left(\frac{2IC_r 10^{IID_r/20}}{IC_r^{IID_r/10} - 1}\right).$$

In the decoder the following reconstruction is done:

$$\hat{L}[k] = \beta L_O[k] + (1-\gamma) R_O[k]$$
$$\hat{R}[k] = (\beta - 1) L_O[k] + \gamma R_O[k]$$
$$\hat{C}[k] = (1-\beta) L_O[k] + (1-\gamma) R_O[k].$$

where $\hat{L}[k]$ is an estimate of $L[k]$, $\hat{R}[k]$ an estimate of $R[k]$ and $\hat{C}[k]$ an estimate of $Cs[k]$. The parameters β and γ are determined in the encoder and transmitted to the decoder.

Knowing all this, the functions that are used for the post-processing are:

$$w_l = f_1(\alpha_l) f_2(\beta)$$
$$w_r = f_3(\alpha_r) f_4(\gamma),$$

here $f_1 \ldots\ldots f_4$ can be any function. For example to apply stereo widening on the rear channels:

$$f_1(\alpha) = f_3(\alpha) = \sin(\alpha)$$
$$f_2(\beta) = f_4(\beta) = \begin{cases} \sin(0.5\pi\beta) & if \quad 0 < \beta < 1 \\ 1 & if \quad \beta \geq 1 \\ 0 & if \quad \beta \leq 0. \end{cases}$$

For the filter functions $H_1 \ldots\ldots H_4$ the following functions are then chosen (in the z-domain):

$$H_1(z) = H_4(z) = 0.8(1.0 + 0.2z^{-1} + 0.2z^{-2})$$
$$H_2(z) = H_3(z) = 0.8(-1.0z^{-1} - 0.2z^{-2}).$$

This invention can be applied in any multi-channel audio-coder that creates a compatible stereo down-mix. The invention can be applied in two ways:
It can be applied before transmission to provide an improved down-mix for decoders that can only decode the stereo audio.

It can also be applied in decoders that can handle the whole bit-stream as a user-selectable option to make improved stereo reproduction possible.

CLAIMS:

1. An audio encoder as described hereinbefore.

2. An audio decoder as described hereinbefore.

3. An audio encoding method as described hereinbefore.

4. An audio decoding method as described hereinbefore.

5. An audio signal as produced by the method of Claim 3.

6. A storage medium having stored thereon a signal as claimed in Claim 5.

ABSTRACT:

This document gives a technical description of a multi-channel parametric audio coding system. The goal of this system is to describe an m-channel signal by an n-channel signal, with $n<m$, and parameters describing a spatial image in order to reconstruct the m-channel signal.

5

Fig. 1